『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装

著者	小木曽 智信, 中村 壮範
ページ	1-141
発行年	2009-03-24
シリーズ	国立国語研究所内部報告書 ; LR-CCG-08-04
URL	http://doi.org/10.15084/00002847

国立国語研究所内部報告書（LR-CCG-08-04）

『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装

小木曽 智信・中村 壮範

平成21年3月
大規模汎用日本語データベースの構築とその活用に関する調査研究

国立国語研究所内部報告書（LR-CCG-08-04）

『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装

小木曽智信
中村　壮範

平成21年3月
大規模汎用日本語データベースの構築とその活用に関する調査研究

目次

はじめに

本稿は『現代日本語書き言葉均衡コーパス』（BCCWJ）の形態論情報を格納するデータベース（「形態論情報データベース」）の設計と実装について記述したものである。形態論情報データベースは、国立国語研究所（形態論情報サブグループ）において運用を行っており、形態素解析辞書 UniDic の元となる見出し語のデータを格納するとともに、UniDic による解析結果を取り込んでコーパスとして利用することを可能にしている。

UniDic の基本設計は伝康晴氏（千葉大学・特定領域研究「日本語コーパス」電子化辞書班班長）によるものであり、その詳細は伝康晴ほか（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」（『日本語科学』22 号, pp.101-122）に論じられている。

本稿の執筆者等は、この基本設計に拡張を加えつつ階層化された辞書見出しとコーパスを格納するデータベースシステムを実装した。本稿では、このデータベースの設計・実装に関する詳細を述べるとともに、運用に関する基本的な情報をあわせて記述する。「形態論情報データベース」の利用者の手引きとするとともに、短単位を基礎とする新たなデータベース開発の参考資料として利用されることを期待している。

本書で扱うのは専らデータベース上での設計と実装、およびデータベースの利用に関する事柄である。UniDic そのものの基本設計については前掲の伝（2007）を、データの言語単位に関する仕様（短単位・長単位等）については『『現代日本語書き言葉均衡コーパス』形態論情報規程集』（以下『形態論情報規程集』）を、そして形態素解析辞書 UniDic については「UniDic ユーザーズマニュアル」をそれぞれ参照されたい。

なお、本書で記述するデータベースの仕様は 2009 年 2 月時点での状態に基づくものであり、今後変更される可能性がある。

2009 年 2 月 23 日　　小木曽智信・中村壮範

1. 形態論情報データベースの概要

形態論情報データベースの主な利用目的は、次の 3 点である。

1. 形態素解析辞書 UniDic の元となる見出し表・活用表を格納し、見出し語の追加・修正作業を行う
2. BCCWJ の短単位で解析されたテキストを格納し、人手による修正を行ったコアデータを作成する
3. 短単位で解析されたテキストを格納し、コーパスを利用した研究に利用する

1 は辞書見出し、2, 3 はコーパスのデータを扱うことになる。これに対応して、形態論情報データベースは、1 の辞書見出しを格納する「辞書データベース」と 2, 3 のコーパスを格納する「コーパスデータベース」に分かれている。コーパスの形態論情報と辞書の情報を同一に保つ必要があるため、それぞれのデータベースは中間に辞書見出し表から生成される「語彙表」を挟んで連係している。コーパスに出現したすべての語は、原則として語彙表のいずれかのレコードと関連付けられる。

形態素解析辞書の作成という観点から見たときには、1, 2 は形態素解析辞書 UniDic の元となるデータを用意するための作業である。1 の見出し表を組み合わせることにより解析辞書の見出し表（辞書）が生成され、2 のコアデータから学習用コーパスが作られる。この二つのデータ元に、機械学習により形態素解析辞書が作成される。

3 はこの形態素解析辞書によって解析されたテキストデータを学習コーパスと同様の形式で格納したものである。このデータは言語研究に利用するだけではなく、辞書の整備（未登録の語を見つけ出し追加する等）のためにも利用される。

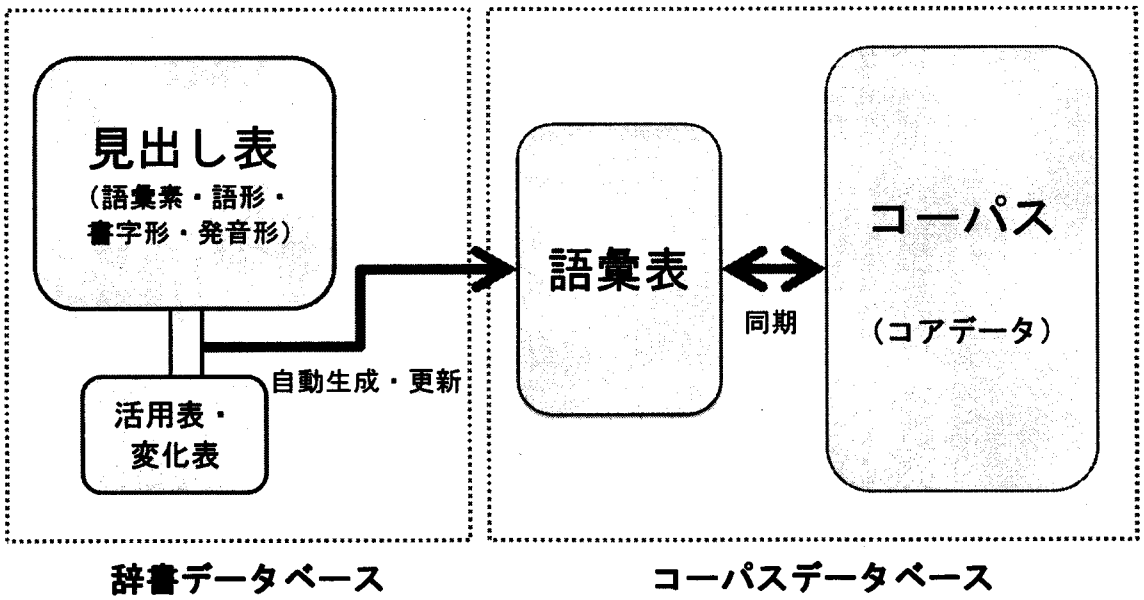


図 1　形態論情報データベース全体図

2. データベースシステム

2.1.データベースシステムの概要

「形態論情報データベース」は、データベースソフト（DBMS）に Microsoft SQL Server を、クライアントに Microsoft Access で作成した専用アプリケーションを用いるクライアント・サーバ型のシステムとして構築されている。以下では、このシステムのネットワーク構成、ソフトウェア（サーバ及びクライアント）、サーバのハードウェアについて概略を説明する。最後に、このシステムの長所と短所について簡単に述べる。

2.2.ネットワーク

形態論情報サブグループでは、クライアントマシンとユーザの管理のために Windows ドメインを導入しており、このドメイン中に SQL サーバを置いている。ドメインはドメインコントローラのほか、クライアントマシン（Windows XP，一部 Vista）約 20 台、SAMBA サーバ（形態素解析辞書学習用ワークステーション）で構成されている（図 2）。LAN 回線はギガビットイーサネットである。図には示していないが、実際にはドメインコントローラ・SQL サーバのバックアップ用のマシンが常時稼働している。

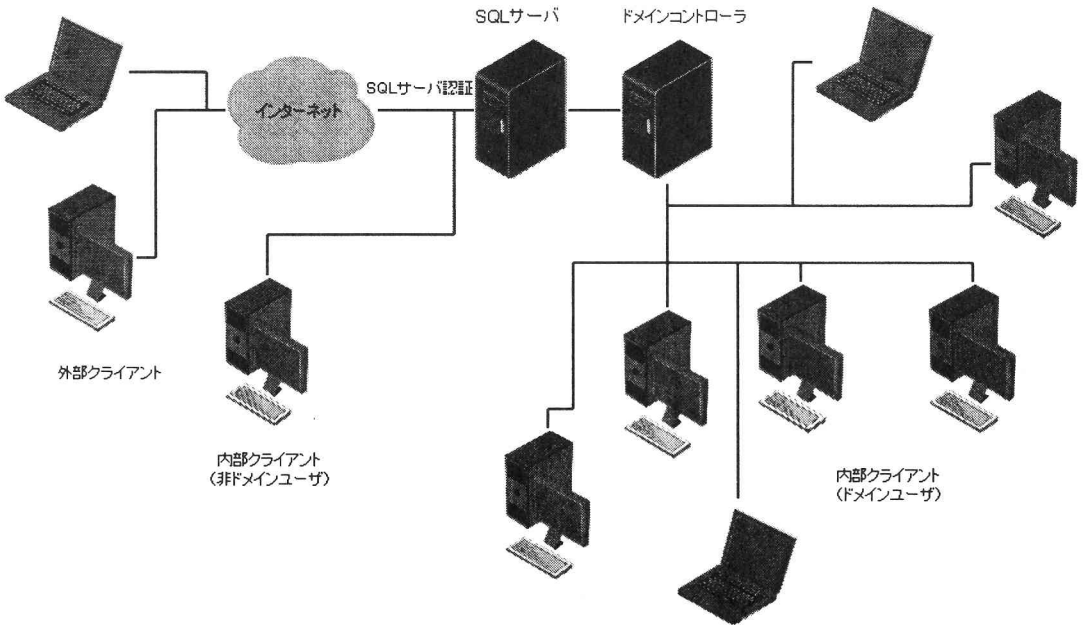


図 2　形態論情報データベースのサーバとクライアント

SQL サーバのユーザ認証は混在モードとし、ドメインによるユーザ認証（Windows 認証）と、SQL サーバ認証の両方に対応している。ドメインユーザは Windows 認証により、ドメイン外のマシンからのアクセスは SQL サーバ認証による。

所外からのアクセスについては、VPN（passportOne）によってインターネット越しの接続を可能にしている。この場合はすべて SQL サーバ認証となる。

2.3.データベースサーバ

サーバ OS には Windows 2003 Server R2 Standard x64 Edition、データベース管理システム（DBMS）として Microsoft SQL Server 2005 Standard Edition（SP2）を利用している。十分なメモリを利用するためいずれも 64 ビット版（x64 Edition）を利用している。

ハードウェアのスペックは次の通りである。

メモリ：24.0GB
CPU ：Intel Xeon X5355 ×2
HDD：1.0TB（RAID5）

SQL Server の規定の照合順序（COLLATE）は Japanese90Bin2 としている。これは BCCWJ で用いられる規定される文字（JIS X 0213 の文字集合）を適切に扱えるようにするためである。

なお、オリジナル関数・ストアドプロシージャ・テーブルなど全てのデータベース上のオブジェクトには、SQL Server の「拡張プロパティ」によって説明が付けられている。

2.4.クライアントアプリケーション

クライアントアプリケーションは Microsoft Access で開発した。一般に小規模データベースで用いる mdb 形式や accdb 形式ではなく、データを全てサーバに置き Access はクライアントとしての機能だけを果たす adp 形式で作成している。Access のバージョンは 2000 以降に対応している。クライアントマシンには原則として Access のインストールが必要であるが、無償配布されている Access ランタイムを用いることにより、Access がインストールされていないクライアントからでも利用可能である。

Access 標準の機能を用いることにより、エンドユーザが作業に必要なクエリ（ビュー）を GUI で作成して作業に用いることも可能となっている。

クライアントアプリケーションの詳細については、5 辞書データベース用アプリケーション、6 コーパスデータベース用アプリケーション・大納言、7 Web アプリケーション・中納言 を参照されたい。

2.5.システムの性能と評価

2.5.1.規模と処理速度

2009年2月現在、形態論情報データベースに格納されたデータの規模は次の通りである。

表 1　形態論情報データベースの規模

データベース	レコード数
辞書データベース	約21万語（書字形）
語彙表	約80万語
コーパスデータベース	約1.8億語※

※BCCWJ以外のデータや重複分を含む

システムの処理速度を示す参考値として、この状況下においてコーパスデータベース用アプリケーション「大納言」を使用して検索を行った際の処理速度をまとめた。いずれも実作業で多く発生する処理である。実際の検索速度は条件によって大きく異なる場合がある。

表 2　コーパスの検索速度（例）

検索の種類	検索対象コーパス	ヒット件数	所要時間
短単位検索　（出現書字形「国語」を完全一致で検索）	約20万語	12	1秒以下
	約200万語	44	1秒以下
	1億8千万語	2746	1秒以下
全文検索（「日本人なら」を検索）	約20万語	1	1秒以下
	約200万語	4	1秒以下
	1億8千万語	117	約13秒
高度な検索（前後の三品詞を組み合わせた検索）	約20万語	2	約2秒
	約200万語	14	約3秒
サンプルID検索（PB10_00047）	約20万語	1243	1秒以下
	約200万語	1243	1秒以下

※　全文検索はSQL Server 2005標準の機能によるものである。

※　サンプルID検索は検索対象コーパスを増やしてもコストは変わらない。

辞書データベースの側では、見出し語の辞書登録に際してリアルタイムで見出し語展開までを行っているが、これも1秒以内に完了し、作業に支障はない。

データベースの同時接続ユーザは20名ほどであるが、排他処理を含め問題は生じていない。

2.5.2.開発コストとライセンス

システムを短期間で開発して実用に供する必要があったことから、アプリケーションの作成が比較的容易であり、一般の会社等での利用事例が多い Microsoft SQL Server と Access の組み合わせを採用した。これにより、実際に数ヶ月という短期間で実用的なシステムが構築できたのみならず、その後も作業者の要望にあわせた作り込みが可能となった。多くのユーザにとって以前から使い慣れた環境で作業できるため、余計な教育コストが掛からない点も長所といえる。DBMS が提供する管理ツール（Microsoft SQL Server Management Studio）についても、使い勝手がよく習熟が容易であった。

一方、商用ソフトウェアであるため、サーバ・クライアントの双方にライセンスが必要である。費用の点のみであれば、開発・メンテナンスに要するコストの低減と比較すれば、導入コストについては十分に元が取れていると考えられる。しかし、作成したソフトウェアをシステムごと配布するような自由な利用が難しくなっている（無償の機能制限版 Express Edition を用いることにより配布自体は可能である）。

3. 辞書データベース

3.1. 辞書データベースの概要

辞書データベースは、形態素解析辞書 UniDic の元となる見出し語のデータベースである。見出し語のテーブルのほか、活用表などの辞書作成に必要な情報からなる。

辞書データベースの基本となる見出し表は、UniDic の見出し設計にあわせて作成された「短単位語彙素」、「短単位語形」「短単位書字形」「短単位発音形」の 4 つである。UniDic では次のような階層化された見出し語が設定されている*。

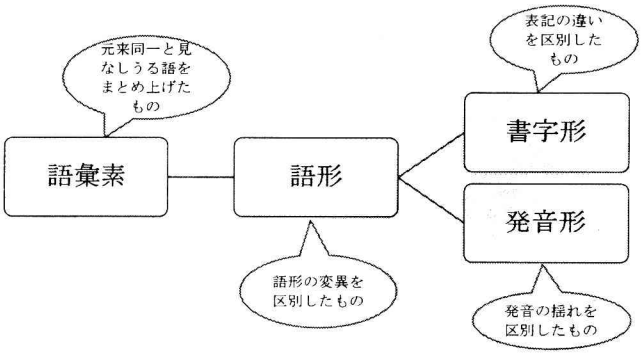


図 3　UniDic の見出し設計

「語彙素」は国語辞典の見出し語に相当するレベルで、語の意味や語の出自などの情報はここに記述される。

「語形」は異語形を区別するレベルで、たとえば「アマリ（余り）」に対する「アンマリ」「アンマシ」「アンマ」といった異語形、上一段活用と文語上二段活用といった活用の違いのほか、可能動詞形もここで区別される。

「書字形」は異表記を区別するレベルで、漢字を使うか仮名書きするかといった違いのほか、送り仮名の揺れもここに記述される。

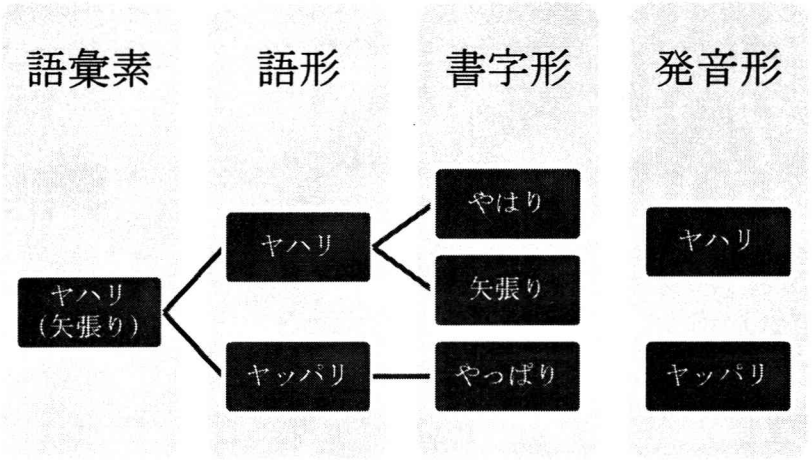
「発音形」は発音やアクセントなどの情報が記述される。

辞書データベースの見出し表はこの階層をそのまま反映している。各テーブルの詳細については 3.2 で述べる。

図 4　辞書データベース短単位表のテーブル設計

*伝康晴ほか（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22 号，pp.101-122）

各見出し語は、具体的には次のように階層化された形で格納されることになる。



※発音形は語形から直接結合する

図 5　UniDic の見出し構造の例

辞書データベースには、見出し表のほかに、活用語を展開するための「活用表」と「活用型表」「活用形表」、語頭変化形を展開するための「語頭変化表」、語末変化形を展開するための「語末変化表」が存在する。

短単位語形は、語頭変化・語末変化・活用のそれぞれの変化をこの順で反映して展開される。語頭・語末変化については 3.4 で、活用の詳細については 3.5 で、出現形展開処理の全体については 3.6 で述べる。

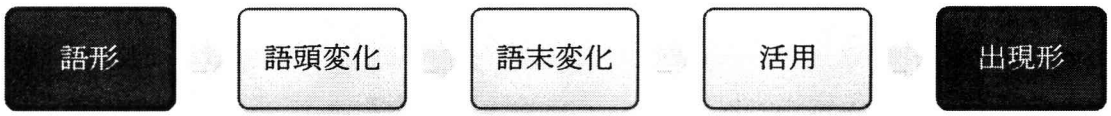


図 6　出現形展開の流れ

データベース上では、各階層の見出し表のレコードはユニークな ID によって関連付けられており、各 ID は計算によって階層関係が確認できるように設計されている。また、見出し表の間では、レコードの生成や削除に関連する制約が付けられている。この ID の計算方法と見出し表の間の制約については、3.7 で述べる。

見出し表に準ずるものとして、「書字形構成漢字テーブル」がある。これは、漢字の使用頻度をコーパス中で使用された語ごとに数えることを可能にするためのテーブルで、書字形テーブルと「漢字テーブル」に関連付けられている。漢字テーブルは漢字の音訓や学年配当など、漢字そのものに関する情報を格納した表である。書字形構成漢字テーブルについては 3.8 で述べる。

このほかに、見出し語入力のための各種情報や、コーパスから取得した頻度等を格納するテーブルが存在する。これらの詳細は、3.9 で述べる。

3.2.見出し表

3.2.1.見出し表の概要

3.1 で見たとおり、見出し表は４つの階層が ID で関連付けられて構成されている。各見出し表の列名と、見出し表の間の関連付けを図 7 に示す。

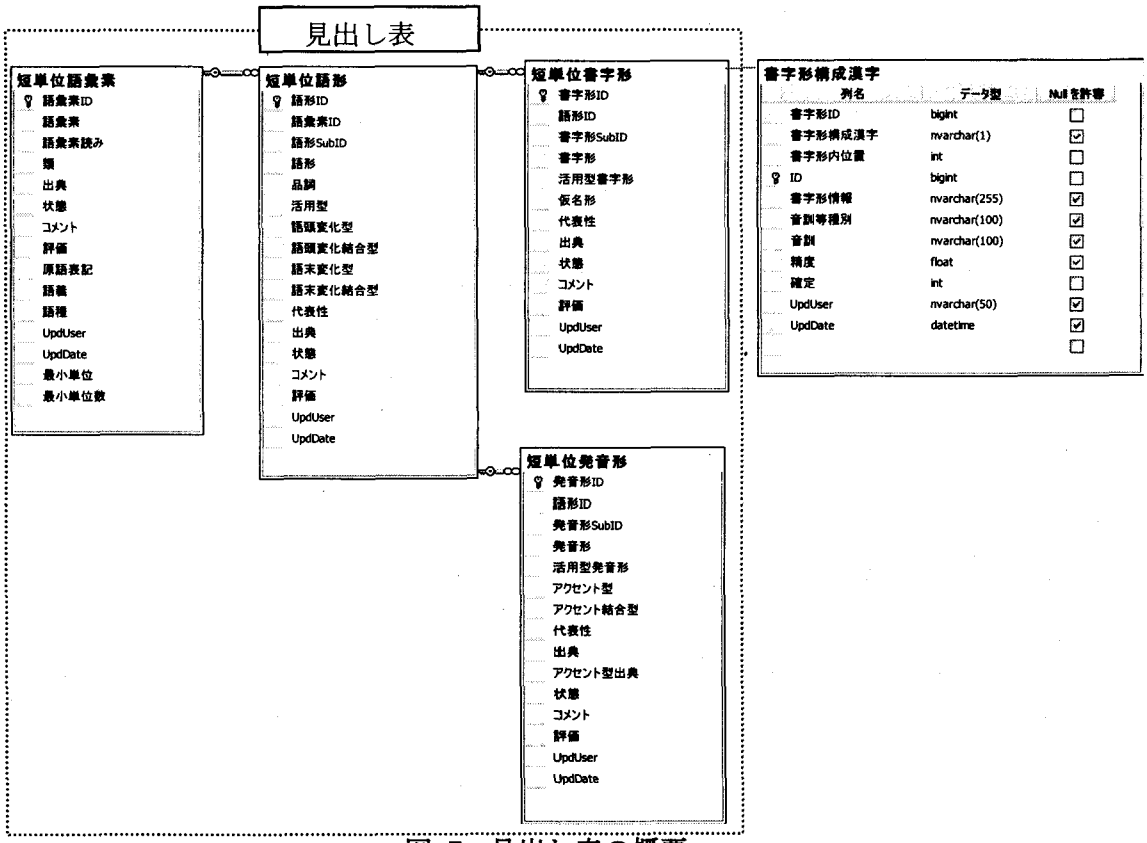


図 7　見出し表の概要

以下では、特に重要な短単位語彙素テーブルから短単位発音形テーブルまでの短単位見出し表について説明する。見出し表共通の属性については 3.2.6 でまとめて説明する。また、3.3 で、各テーブル更新時に自動実行される処理（トリガ）について説明する。短単位書字形テーブルと関連付けられる書字形構成漢字テーブルについては 3.8 で述べる。

なお、見出し表に記載されるのは原則として基本形（終止形）のみであり、各活用形・濁音形などは、活用表・変化表によって生成される。これらの表と展開処理については 3.5・3.6 で別途説明する。また、各表を関連付ける ID の計算方法については 3.7.2 で説明する。

3.2.2.短単位語彙素テーブル

短単位語彙素テーブルには表 3 の情報が格納される。

表 3　短単位語彙素テーブルの列

Index	入力	列名	説明
◎	自動	語彙素 ID	主キー（連番）
○	必須	語彙素	辞書見出しの代表表記に相当（漢字仮名混じり表記）
○	必須	語彙素読み	辞書見出しに相当（カタカナ表記）
○	※	語彙素細分類	語彙素を語義等によって更に細分する
○	必須	類	見出し語の類（体・用・相）等による区別（品詞の上位概念に相当）
	必須	語種	見出し語の出自による区別
	自動	最小単位	見出し語を最小単位に分割した場合の数
		原語表記	（語彙素細分類に統合、廃止）
		出典	共通属性
		コメント	共通属性
		状態	共通属性
		評価	共通属性
	自動	更新日時	共通属性
	自動	更新ユーザ名	共通属性

◎：主キー、○：一意のクラスタ化インデックス

- 「語彙素 ID」はユニークな主キーで、1 からの連番である。ただし、見出し語の削除によって間隔が開いている場合がある。短単位語形テーブルとの関連付けはこの ID による。
- 「語彙素」「語彙素読み」「類」「語種」は入力が必須である。「語彙素読み」を持たない補助記号類については空文字列を入力する（null は許容されない）。
- 「語彙素細分類」は語彙素を語義や語源によって更に細かく区別する場合の値で、通常は空文字列である。ライト-right, ライト-light のように、「語彙素」「語彙素読み」「類」「語種」の 4 属性では区別ができない場合に入力が必須となる。
- 「語種」は表 4 の 7 種類のいずれかである。このうち、固有名、記号については、入力された「類」によって一意に決められる。そのため辞書登録ツールでは自動入力される。「※」は作業用の値で、見出し入力時に語種が不明で、調査が未了であ

ることを示す。一方「不明」は、調査の結果、複数の語源説があるなどして語種不明であることが判明したことを示す。

表 4　語種の値

値	説明
和	和語
漢	漢語
外	外来語
混	混種語
固	固有名
記号	記号
不明	語種不明
※	確認中

- 「最小単位」は、短単位語彙素の新規登録時にトリガによって「語彙素読み」と同一の文字列が入力される。和語・混種語・語種不明の場合には、これに次例のような書式で最小単位境界を作業者が記入する。

  「アシ/」（足）

  「アシ/アト」（足跡）

  「ジュウ/バコ」（重箱）

  すなわち、１最小単位から成る場合には末尾に「/」を追加し、２最小単位以上から成る場合には単位の境界に「/」を入力する（したがって、和語・混種語・語種不明でありながら最小単位に「/」を含まないものは未処理であることを示す）。

  なお、漢語・外来語・固有名・記号の場合には、最小単位数は容易に計算ができるため入力を要しない。すなわち、短単位の定義から外来語・固有名・記号は常に１最小単位であり、漢語の場合は代表表記の漢字の文字数分である。
- テーブルに付与された制約（クラスタ化インデックス・語彙素 uniq）により、同一の「語彙素」「語彙素読み」「語彙素細分類」「類」を持つエントリの重複は許されない。したがってこの５属性の組み合わせによって短単位語彙素テーブル中のエントリが一意に決まる。よって、短単位語彙素の同定には「語彙素 ID」または「語彙素」「語彙素読み」「語彙素細分類」「類」のセットのいずれかを用いることができる。
- 短単位語彙素テーブルのレコードを削除する場合には、必ず子や孫となる語形・書字形・発音形を先に削除しておかなければならない（ツールでは子や孫となる見出し語ごと削除することができるが、データベース上ではカスケード削除には設定していない）。

3.2.3.短単位語形テーブル

短単位語形テーブルには表 5 の情報が格納される。

表 5　短単位語形テーブルの列

Index	入力	列名	説明
◎	自動	語形 ID	主キー
	自動	語彙素 ID	親の語彙素の ID
	自動	語形 SubID	同一語彙素に関連付けられる語形の連番
○	必須	語形	異語形を区別するレベルの見出し（カタカナ）
	必須	品詞	品詞
	※	活用型	辞書登録活用型 ※活用語の場合は必須
		語頭変化型	濁音化などの語頭音変化の種類（型）
		語頭変化結合型	後続要素の語頭変化形への制約の種類（型）
		語末変化型	促音化などの語末音変化の種類（型）
		語末変化結合型	前接要素の語末変化形への制約の種類（型）
		代表性	共通属性
		状態	共通属性
		評価	共通属性
	自動	更新日時	共通属性
	自動	更新ユーザ名	共通属性

- 短単位語形テーブルの新規レコードを入力するには、必ず親となる語彙素が入力済みでなければならない。また、短単位語形テーブルのレコードを削除する場合には、必ずこの見出し語の子となっている書字形・発音形を先に削除しておかなければならない（ツールでは子の見出し語を自動削除することができるが、データベース上ではカスケード削除には設定していない）。
- 「語形 ID」は短単位語形テーブルの主キーで、語彙素 ID に一定数をかけて語形 SubID を足したもの。「語彙素 ID」は当該語形の親となる語彙素の ID。「語形 SubID」は同一語彙素の元にぶらさがる語形にふった 1 からの連番。ID 生成の詳細は 3.7.2 を参照のこと。ツールにおいて語形 ID の入力は自動で行われる。
- 「語形」「品詞」は入力が必須である。また、活用語の場合には「活用型」も入力が必須である。
- 「語形」には、たとえば語彙素「やはり」の場合、「ヤハリ」の異語形である「ヤッパリ」「ヤッパシ」「ヤッパ」などがぶら下がることになる。なお、語頭が濁音になる形は後述の語頭変化型で生成するため個別には入力しない。

動詞の場合には、文語形、可能動詞形についてもこのレベルで区別する。したがって語彙素「書く」の語形として、五段活用動詞（五段-カ行-一般）「カク」のほかに、下一段活用（下一段-カ行）の「カケル」、四段活用動詞（文語四段-カ行）の「カク」がぶら下がることになる。

- 「品詞」には、当該語の品詞として適切なものを選択して入力する。選択可能な品詞は、資料①の品詞一覧を参照。なお、選択可能な品詞は当該語形の親となる語彙素の「類」によって制限される。そのため、ツールでの入力時には選択肢が自動で絞られる。
- 「活用型」には、当該語が活用語である場合に限り、活用型を選択する。活用型は品詞によって選択できる型が変わるため、ツールでの登録時には選択肢が自動で絞られる。この活用型は辞書登録用の活用型であり、コーパス中の語の活用型とは一部違いがある（特に区別する必要がある場合には辞書登録活用型と呼ぶ）。活用型は、資料②の活用型一覧を参照。
- 「語頭変化型」は濁音化などの語頭音変化の種類を示す。たとえば「カイ（貝）」の場合、ここに「カ濁」型を指定することにより、基本形「カイ」と濁音形「ガイ」の二つの語形が生成されることになる。変化形を持たない語の場合は指定しない。詳細は 3.4.2 を参照。語頭変化型の種類は資料④（105 ページ）参照。
- 「語末変化型」は促音化などの語末音変化の種類を示す。たとえば「サンカク（三角）」の場合、ここに「ク促」型を指定することにより、基本形「サンカク」と促音形「サンカッ」の二つの語形が生成されることになる。変化形を持たない語の場合は指定しない。詳細は 3.4.3 を参照。語末変化型の種類は資料⑤（106 ページ）参照。
- なお、特定の活用形の自動生成されない書字形を登録したい場合には、その基本形を語形として入力し、特殊な活用型を指定する。詳細は 3.5.7 を参照。

3.2.4.短単位書字形テーブル

短単位書字形テーブルには表 6 の情報が格納される。

表 6 短単位書字形テーブルの列

Index	入力	列名	説明
◎	自動	書字形 ID	主キー
	自動	語形 ID	親となる語形の ID
	自動	書字形 SubID	同一語形に関連付けられる書字形の連番
○	必須	書字形	表記を区別するレベルの見出し
	必須	仮名形	書字形をカタカナ表記にしたもの
	自動	活用型書字形	（関数で生成）
		代表性	共通属性
		状態	共通属性
		評価	共通属性
	自動	更新日時	共通属性
	自動	更新ユーザ名	共通属性

- 短単位書字形テーブルの新規レコードを入力するには、必ず親となる語形が入力済みでなければならない。短単位書字形テーブルのレコードを削除した場合には、関連付けられる書字形構成漢字のレコードがトリガによって削除される。
- 「書字形 ID」は短単位書字形テーブルの主キーで、語形 ID に一定数をかけて書字形 SubID を足したもの。「語形 ID」は当該書字形の親となる語形の ID。「書字形 SubID」は同一語形の元にぶらさがる書字形にふった 1 からの連番。ID 生成の詳細は 3.7.2 を参照のこと。ツールにおいて書字形 ID の入力は自動で行われる。
- 「書字形」は当該語の表記を記述する。活用語の場合には原則として活用語尾が仮名書きで含まれなければならない。
- 「仮名形」は当該語をカタカナ表記にしたもの（日本語入力辞書への応用を考慮したもので、形態素解析には利用しない）。
- 「活用型書字形」はデータベース内部における活用形展開に必要な書字形に関する情報である。たとえば形容詞「赤い」のウ音便は、漢字表記の場合には「赤う」と末尾のみ変化させればよいが、かな書きされる「あかい」の場合には「あこう」と二文字分変化させる必要がある。このため、内部の活用型では「形容詞-カイ+一般」と「形容詞-カイ+かい」とに区別されている。このときの「+」以降の部分が活用型書字形である。現在のデータベースでは、この情報を静的に格納せず、データベー

ス上の関数によって活用型と書字形から動的に生成している。この関数については資料⑧オリジナル関数一覧を参照。

- なお、当該書字形の親となる語形が特殊な活用型である場合には、「書字形」「仮名形」には基本形ではなく活用後の形（出現形）を入力する。

3.2.5.短単位発音形テーブル

短単位発音形テーブルには表 7 の情報が格納される。

表 7　短単位発音形テーブルの列

Index	入力	列名	説明
◎	自動	発音形 ID	主キー
	自動	語形 ID	親となる語形の ID
	自動	発音形 SubID	同一語形に関連付けられる発音形の連番
○	必須	発音形	発音を区別するレベルの見出し
		アクセント型	アクセント型（アクセント核のある位置）
		アクセント修飾型	活用によるアクセント変化の種類（型）
		アクセント結合型	前接（後続）要素との結合時のアクセント変化の種類（型）
	自動	活用型発音形	（関数で生成）
		代表性	共通属性
		状態	共通属性
		評価	共通属性
	自動	更新日時	共通属性
	自動	更新ユーザ名	共通属性

- 短単位発音形テーブルの新規レコードを入力するには、必ず親となる語形が入力済みでなければならない。
- 「発音形 ID」は短単位発音形テーブルの主キーで、語形 ID に一定数をかけて発音形 SubID を足したもの。「語形 ID」は当該発音形の親となる語形の ID。「発音形 SubID」は同一語形の元にぶらさがる発音形にふった 1 からの連番。ID 生成の詳細は 3.7.2 を参照のこと。ツールにおいて発音形 ID の入力は自動で行われる。
- 「発音形」は当該語の発音をカタカナで記述する。発音を示すものであるため助詞「は」なども「ワ」で表される。長音は常に「ー」で、また「ヅ」「ヂ」は常に「ズ」「ジ」で表される。

- 「アクセント型」は当該語のアクセントをアクセント核の位置を示す数字で表す。すなわち、頭高型は「1」、平板型は「0」となる。
- 「アクセント修飾型」は特定の活用形を取る場合に起こるアクセント型の変化の種類を記述する。詳細は UniDic ユーザーズマニュアルを参照。
- 「アクセント結合型」は複合語を作ったり、自立語に付属語が結合したりする際に起こるアクセント型の変化の種類を記述する。詳細は UniDic ユーザーズマニュアルを参照。
- 「活用型発音形」はデータベース内部における活用形展開に必要な発音形に関する情報である。たとえばカ行五段活用動詞のイ音便の発音形は、「書く」の場合には「カイ」と「イ」になるが、「聞く」のように語幹がイ段（またはエ段）で終わる場合には「キー」と長音符号に置き換える必要がある。このため、内部の活用型では「五段-カ行-一般」を「=一般」と「=イエ段」とに区別している。このときの「=」以降の部分が活用型発音形である。現在のデータベースでは、この情報を静的に格納せず、データベース上の関数によって活用型と発音形から動的に生成している。この関数については資料⑧オリジナル関数一覧を参照。

3.2.6.見出し表の共通属性

見出し表（短単位語彙素テーブル、短単位語形テーブル、短単位書字形テーブル、短単位発音形テーブル）に共通して付けられるレコードに関する情報がある。これらを表 8 に示す。主として見出し語に関するメタ的な情報や管理情報であり、必ずしも必須の情報ではない。

表 8　見出し表の共通属性

列名	説明
出典	当該の見出し語のソースとなった資料
状態	当該の見出し語の利用の状態を示す
代表性	当該見出し語が同階層において代表性を持つかどうか（未整備）
コメント	当該の見出し語に関する情報（自由記述）
評価	児童向けの表記、創作固有名詞等の情報
更新日時	最終更新日時
更新ユーザ名	最終更新ユーザ

- 「出典」は当該の見出し語のソースとなった資料を示す。最初に登録された時点での出典を示すもので、コーパスの追加によって他のソースでの使用が確認された場合には更新されるわけではない。「出典」の種類については資料⑥参照。

- 「状態」は当該見出し語の形態素解析辞書での利用状態を表すもので、1文字の記号（及びその組み合わせ）で示す。たとえば「仮」は仮登録であることを示し、確認が完了するまで形態素解析辞書には出力されない。また「Z」はコアデータに出現したことから辞書登録を行ったものの、特殊な語であるため形態素解析辞書には出力しないことを示す。その他の「状態」の一覧は資料⑦を参照。
  なお「状態」属性は、短単位語彙素・語形・書字形・発音形の全ての階層に付与することができるが、実際の解析辞書作成用データの出力に当たっては短単位書字形テーブルの状態だけが参照される。
- 「代表性」は、当該見出し語が、同じ階層のグループの中で代表となることを示すもので、真偽値（True／False）で表される。たとえば語形「ヤハリ」「ヤッパリ」「ヤッパシ」「ヤッパ」のなかで「ヤハリ」を代表形とする場合に「ヤハリ」を代表性 True とすることになる。ただし、現在は完全な運用を行っていない（その階層のグループの中で最初に作られたものが代表性を持つように自動処理されている）。

3.3. 見出し表のトリガ

4つの見出し表は、レコードの新規登録時や更新時にデータベース上で既定の自動処理が実行される（トリガによる処理）。各見出し表のトリガで行われる処理には「語彙表生成処理」「更新情報記入処理」「書字形構成漢字処理」の三つがある。

「語彙表生成処理」は、辞書データベースとコーパスデータベースをつなぐ語彙表に、見出しを追加したり、更新したり、削除したりするものである。処理の内容を表 9 に示した。見出し語を新規登録する場合には、短単位語彙素・短単位語形・短単位書字形・短単位発音形の4つの見出しテーブルがそろったときに初めて語彙表の見出し生成が実行される。見出し表のレコードの削除時には、対応する語彙表の見出しも削除される。語彙表生成の詳細については 3.4～3.6 を参照のこと。

表 9　語彙表生成処理

対象テーブル	短単位語彙素テーブル・短単位語形テーブル・短単位書字形テーブル・短単位発音形テーブル
実行条件	新規登録時に語彙素から書字形・発音形までがそろい語彙表生成が可能になったとき、または次の8属性（語彙素・語彙素読み・語彙素細分類・語形・品詞・活用型・書字形・発音形）のアップデート時
処理内容	語彙表の見出しを追加・更新・削除する。

「更新情報記入処理」はその見出し語を更新した日時とユーザ名を、各見出し表の「更新日時」「更新ユーザ名」に記入する処理である。処理の内容を表 10 に示した。

「語彙表生成処理」と「更新情報記入処理」は、見出し表の更新の中でも、語彙表を更新する必要がある重要な情報が更新された場合にのみ実際の処理が行われる。「出典」の修正やコメントの追加などでは語彙表再生成が行われないので、更新情報もアップデートされない。

表 10　更新情報記入処理

対象テーブル	短単位語彙素テーブル・短単位語形テーブル・短単位書字形テーブル・短単位発音形テーブル
実行条件	新規登録時、または次の 8 属性（語彙素・語彙素読み・語彙素細分類・語形・品詞・活用型・書字形・発音形）のアップデート時
処理内容	更新した日時とユーザ名を、各見出し表の「更新日時」「更新ユーザ名」に記入する。語形が削除された場合は語形削除ログテーブルに新規レコードを作成する。

「書字形構成漢字処理」は、短単位書字形に変更があったときに当該書字形に関連付けられている書字形構成漢字テーブルを更新するものである。処理内容を表 11 に示す。書字形構成漢字詳細については 3.8 を参照のこと。

表 11　書字形構成漢字処理

対象テーブル	短単位書字形テーブル
実行条件	新規登録時、または「書字形」のアップデート時
処理内容	書字形構成漢字テーブルに当該書字形に含まれる漢字を追加・更新・削除する。

3.4. 語頭・語末変化

3.4.1. 語頭・語末変化の概要

語頭・語末変化は、連濁などの規則的な現象によって生じる語形変化を反映させた形を生成するための処理である。濁音化などの「語頭変化」と促音化などの「語末変化」に分かれる。特に数詞は複雑な語形変化を起こす。語頭変化と語末変化の両方を起こす語はいまのところ数詞のみである。

3.4.2.語頭変化

語頭変化とは、「語形」が持つ「語頭変化型」に応じて、語形変化による語形を展開する処理である。ここでは、「カ濁」型の語頭変化型を持つ語形「カメ（亀）」を例に説明する。

語頭変化表（資料④）によれば、「カ濁」型には、語頭語形「カ」の基本形と、語頭語形「ガ」の濁音形がある。これにより、語形「カメ」は元の形である基本形「カメ」と、語頭文字を置き換えた濁音形「ガメ」に展開される。基本形と濁音形は語頭変化形 SubID にもとづき違う ID が与えられる。

書字形のレベルでは、濁音形の書字形は、漢字表記の場合には基本形と同じものが使われるが、ひらがな・カタカナで書かれている場合には書字形の先頭部分も変化させたものが出力される。この処理はデータベース上のオリジナル関数とストアドプロシージャによって行われる。

図 8 はこの処理を図示したものである。このうち、辞書データベースに直接登録されているのは語彙素と語形の基本形にあたる部分、及びその配下にある書字形であって、濁音形以下の部分は語頭変化型にもとづき自動で生成されたものである。

なお、語頭変化型の一覧は資料④を参照のこと。語頭変化の種類によっては、半濁音形をもつなど、２つ以上の変化形を持つこともある。

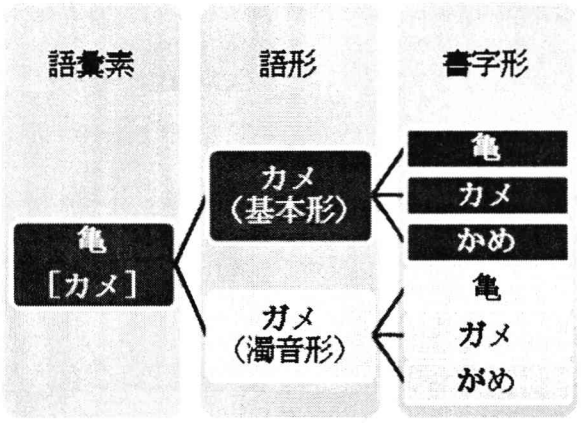


図 8　語頭変化

3.4.3.語末変化

語末変化とは、「語形」が持つ「語末変化型」に応じて、語形変化による語形を展開する処理である。ここでは、「ク促」型の語末変化型を持つ語形「サンカク（三角）」を例に説明する。

語末変化表（資料⑤）によれば、「ク促」型には、語末語形「ク」の基本形と、語末語形「ッ」の促音形がある。これにより、語形「サンカク」は元の形である基本形「サンカク」と、語末文字を置き換えた促音形「サンカッ」に展開される。基本形と促音形は語末変化形 SubID にもとづき違う ID が与えられる。.

書字形のレベルで、促音形の書字形は、漢字表記の場合には基本形と同じものが使われるが、ひらがな・カタカナで書かれている場合には書字形の語末部分を変化させたものが出力される。この処理はデータベース上のオリジナル関数とストアドプロシージャによって行われる。

図 9　語末変化はこの処理を図示したものである。このうち、辞書データベースに直接登録されているのは語彙素と語形の基本形にあたる部分、及びその配下にある書字形であって、促音形以下の部分は語末変化型にもとづき自動で生成されたものである。

なお、語末変化型の一覧は資料⑤を参照のこと。語末変化の種類によっては、２つ以上の変化形を持つこともある。

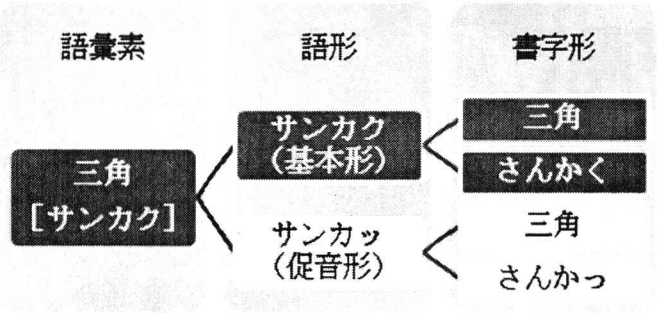


図 9　語末変化

3.5.活用

3.5.1.活用の概要

活用は、語形が持つ活用型に応じて、活用形を展開する処理である。活用型の一覧はデータベースの活用型テーブルに記述されている。活用型の一覧は資料②活用型に、活用形の一覧は資料③に示した。

データベース上では「短単位語形テーブル」と「活用表テーブル」を活用型によって結合することで各活用形を生成する。活用表テーブルは長大になるため、本書では省略したが、表の一部を 3.5.4 で例示した。項目等の詳細については資料⑩を参照のこと。

各活用形の語形（出現形）は、活用表テーブルに記述された活用語尾をもとにして作られる。同様に、その語形の子である書字形・発音形も、活用表テーブルに記述された活用語尾をもとにしてそれぞれの出現形を生成する。

なお、活用語が語頭・語末変化型を持つ場合には、語頭語末変化による語形展開を行った後で活用形が展開される。

3.5.2.活用形の展開

動詞・形容詞等の活用語の場合、短単位語形テーブルに活用型が記述されている。活用表テーブルに接続して、この活用型に応じて各活用形を生成するのが活用形の展開である。

活用に際して、書字形が異なると変化する語尾の部分が異なる場合がある。たとえば、カ行変格活用の動詞「来る」では、仮名で書かれた「くる」の場合、未然形の書字形は「こ」、連用形は「き」だが、漢字で書かれた「来る」では書字形はいずれも「来」である。このように、辞書登録されている書字形によって活用語尾の書字形を変える必要があるため、書字形に「活用型書字形」の情報を持たせて活用形の展開の仕方を変えている。形態論情報データベースでは活用型書字形は関数によって自動で生成するようになっている。

同様の活用語尾変化の違いが、発音形についても起こる。これは主に音便形の処理で発生するもので、たとえば語形が「オイ」で終わる形容詞は、その前がオ段の場合には終止形などの発音形を長音にする必要がある（「トオイ」→「トーイ」）のに対し、それ以外の場合にはその必要がない（「アオイ」→「アオイ」）。このため、発音形に「活用型発音形」の情報を持たせて活用形の展開の仕方を変えている。活用型発音形は関数によって自動で生成するようになっている。

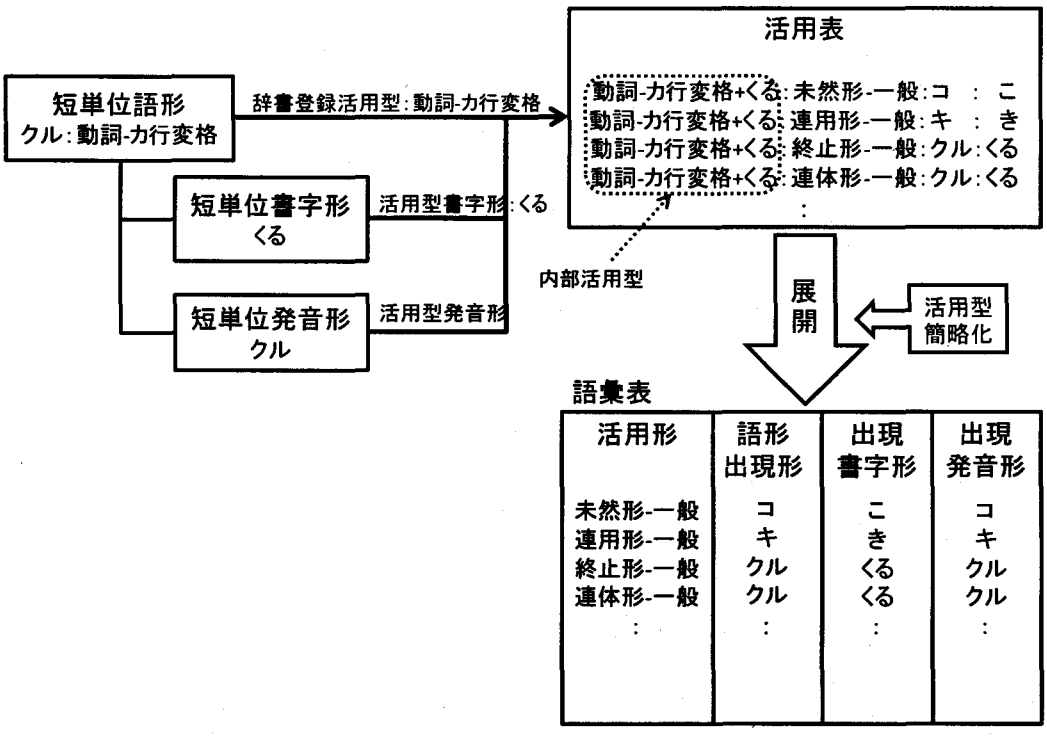


図 10　活用形展開の流れ

このようにして各語形が展開された後、活用型簡略化（3.5.3）が行われ、活用形の展開が完了する。

3.5.3.活用型簡略化

辞書登録活用型に、活用型書字形と活用型発音形を次の書式で付加したものを内部活用型と呼んでいる。活用表は内部活用型を中心にして構成されている。

内部活用型：　辞書登録活用型(+活用型書字形)(=活用型発音形)

活用形の展開が終わった後は、「活用型簡略化テーブル」を使って簡略な活用型に変換している。展開が終われば、詳細な活用型の区別は不要になるためである。コーパスデータベースのデータはこの簡略な活用型で登録されているほか、形態素解析辞書の出力でもこの活用型が使われる。これを単に活用型、ないしコーパス活用型と呼んでいる。

このため、活用型には内部で用いるものを含めて３段階あることになる（表 12）。なかには全く変わらないものもある。

表 12　活用型の例

活用型の段階	例	説明	利用者
辞書登録活用型	カ行変格 形容詞-オイ 下一段-ア行	見出し表（短単位語形）への登録で使われる	見出し追加作業者のみ
内部活用型	カ行変格+くる 形容詞-オイ+一般=オ段 下一段-ア行	データベース内部の処理で使われる	活用表管理者のみ
活用型 （コーパス活用型）	カ行変格 形容詞 下一段-ア行	コーパス・形態素解析辞書の出力で使われる	UniDic の全ユーザ

辞書登録活用型を直接目にすることがあるのは、見出し表（短単位語形）への追加を行う作業者のみである。内部活用型はデータベース内で使われるのみであり活用表を更新する管理者を除き、直接に関わることはない。（コーパス）活用型は、UniDic のエンドユーザを含めた全ての利用者が使うことになる。

なお、活用形展開時には、活用形 ID を与えるために、活用形についても内部活用形が使われている。

3.5.4.活用表

それぞれの活用型がどの活用形を持つかは、辞書データベースの活用表テーブルに記述されている。あわせて3000行を超える膨大な量になるため、本書では省略するが、その一部を以下に例示する。辞書登録型に活用型書字形と活用型発音形の情報を付与した内部活用型とその活用型が持つ活用形を基準としたテーブルになっている。

表 13　活用表の例（カ行変格活用）

内部活用型	活用形	活用語尾	代表性	活用語尾 書字形	活用語尾 発音形	活用語尾 仮名形	アクセント 修飾型	活用形
カ行変格+くる	仮定形-一般	クレ	0	くれ	クレ	クレ		仮定形-一般
カ行変格+くる	仮定形-融合	クリャ	0	くりゃ	クリャ	クリャ		仮定形-融合
カ行変格+くる	命令形	コイ	0	こい	コイ	コイ		命令形-一般
カ行変格+くる	意志推量形	コヨウ	0	こよう	コヨー	コヨウ	M1@1	意志推量形-一般
カ行変格+くる	意志推量形	コヨッ	0	こよっ	コヨッ	コヨッ	M1@1	意志推量形-促音
カ行変格+くる	意志推量形	コヨ	0	こよ	コヨ	コヨ	M1@0	意志推量形-短縮
カ行変格+くる	未然形-一般	コ	0	こ	コ	コ		未然形-一般
カ行変格+くる	終止形-一般	クル	1	くる	クル	クル		終止形-一般
カ行変格+くる	終止形-撥音便	クン	0	くん	クン	クン		終止形-撥音便
カ行変格+くる	連体形-一般	クル	0	くる	クル	クル		連体形-一般
カ行変格+くる	連体形-撥音便	クン	0	くん	クン	クン		連体形-撥音便
カ行変格+くる	連体形-省略	ク	0	く	ク	ク		連体形-省略
カ行変格+くる	連用形-一般	キ	0	き	キ	キ		連用形-一般
カ行変格+来る	仮定形-一般	クレ	0	れ	クレ	クレ		仮定形-一般
カ行変格+来る	仮定形-融合	クリャ	0	りゃ	クリャ	クリャ		仮定形-融合
カ行変格+来る	命令形	コイ	0	い	コイ	コイ		命令形-一般
カ行変格+来る	意志推量形	コヨウ	0	よう	コヨー	コヨウ	M1@1	意志推量形-一般
カ行変格+来る	意志推量形	コヨッ	0	よっ	コヨッ	コヨッ	M1@1	意志推量形-促音
カ行変格+来る	意志推量形	コヨ	0	よ	コヨ	コヨ	M1@0	意志推量形-短縮
カ行変格+来る	未然形-一般	コ	0		コ	コ		未然形-一般
カ行変格+来る	終止形-一般	クル	1	る	クル	クル		終止形-一般
カ行変格+来る	終止形-撥音便	クン	0	ん	クン	クン		終止形-撥音便
カ行変格+来る	連体形-一般	クル	0	る	クル	クル		連体形-一般
カ行変格+来る	連体形-撥音便	クン	0	ん	クン	クン		連体形-撥音便
カ行変格+来る	連体形-省略	ク	0		ク	ク		連体形-省略
カ行変格+来る	連用形-一般	キ	0		キ	キ		連用形-一般

3.5.5.内部活用形と活用形 ID

語彙表の生成にあたって、データベース内部では出現形の差異を反映したさらに詳細な活用形（内部活用形）が用いられる。たとえば、活用型「サ行変格-スル」の命令形では「せよ」「しろ」など複数の形がある。コーパス（形態素解析結果）ではこれらを活用形の名前としては区別しないが、データベース中ではこれに「命令形-一般」「命令形-ロ」のように別の名前・別の ID を与えて区別している。これは語彙表の生成にあたって、実際の書字形や発音形に拠らず、ID のみで語彙エントリをユニークに決定する必要があるためである。語彙表 ID の計算では内部活用形に付与された ID が使われる。内部活用形とその ID は、辞書データベースの活用形テーブルに定義されており、語彙表生成の際に参照される。

3.5.6.活用形テーブルと活用型テーブル

辞書データベースには活用表テーブルの他に「活用形テーブル」と「活用型テーブル」がある。活用形テーブルは活用形 ID の付与に使われ、語彙表の展開に必須である。一方、活用型テーブルは辞書管理ツールで活用型を入力する際に選択するためのデータソースとして利用するものであって、活用形展開時に利用されることはない。「品詞テーブル」も同様である。

3.5.7.特殊な辞書登録活用型

一般の活用表では生成できない特殊な活用形を辞書登録したい場合がある。たとえば、活用語尾までがカタカナ書きされる「イイ（良い）」「デキル（出来る）」や、活用語尾のない特殊な表記「也（助動詞）」、特殊な語形「ま～す」などである。これらをすべて活用表に登録して扱うことは煩雑となるため、現在は特殊な辞書登録活用型を用いて必要な活用形だけを生成できるようにしている。

「無変化○○型:□□」の書式で見出し表に登録することで、□□型のコーパス活用型を持つ、○○形の語形を出力することができる。「無変化○○型」として選択可能なものは活用型テーブルに記載されている。区切り子の「:」は半角である。このような辞書登録活用型をもつ場合、語形には基本形（活用変化前の形）を、その子である書字形・発音形には活用変化後の形を記載する。

例：
	イイ	無変化終止連体型:形容詞
	デキル	無変化終止連体型:上一段-カ行
	也	無変化終止型:文語助動詞-ナリ-断定
	ま～す	無変化終止型:助動詞-マス

この特殊な辞書登録活用型の展開処理は語彙表展開を行うストアドプロシージャの中で行っている。ただし、この方式では基本形を持つ既登録語の一語形として追加することしかできない。また、語形・書字形・発音形の基本形や、アクセント型などの情報を正しく出力することができない。したがって、この方式は暫定的なものであり、今後変更される予定である。

3.6. 語彙表生成のまとめ

語彙表は、語頭・語末変化（3.4 参照）と活用（3.5 参照）を組み合わせて作られる（図 11）。

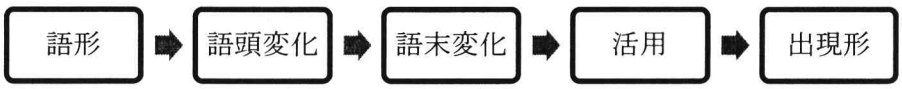


図 11　語彙表生成の流れ

例として「カライ（辛い）」の場合をあげる。「カライ」は、「カ濁」の語頭変化型を持つため、基本形「カライ」と濁音形「ガライ」が展開される。さらに、「カライ」は活用語であるから形容詞の各活用形が展開される。語形の下にある書字形・発音形についても全ての活用形が展開される。図 12 にこの展開の様子の一部を示した。

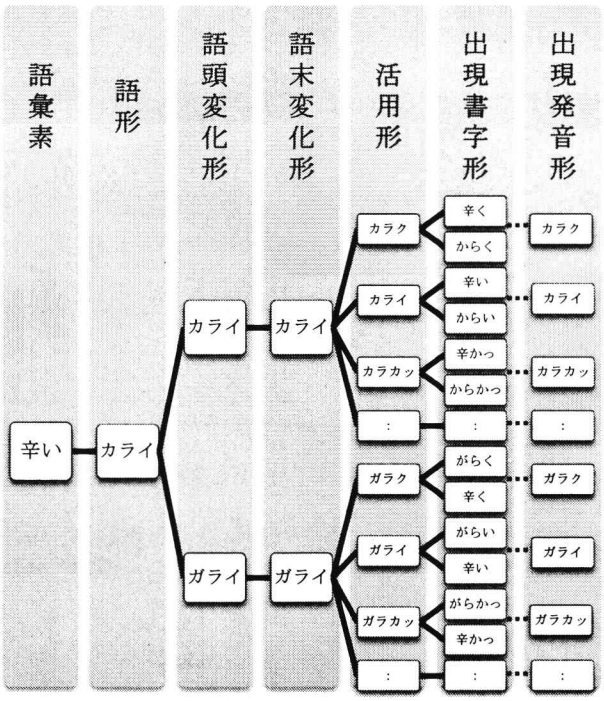


図 12　語彙表生成の例

3.7.見出し表の関連付け

3.7.1.見出し表の関連付けの概要

短単位語彙素・短単位語形・短単位書字形・短単位発音形の４つの見出し表は階層構造を持ち、それぞれの見出し語がIDで関連付けられている。また、４つの見出し全体として重複する値が入力されないようにデータベース上の制約が付けられている。ここでは、この見出し表のIDの計算方法と、見出し表の間の制約について述べる。

3.7.2.見出しID

見出し表はそれぞれのIDによって結合される。各表のIDは親となる見出し語の見出しIDをもとにした計算によりユニークな数字が与えられる。各変化形のIDから親の見出しIDは計算で求めることができる。SubIDは子の階層に位置する見出し語に、親となる見出し語ごとに付与されている1から32（書字形・発音形は256）までの数字（連番）である。

語形ID = 語彙素ID*32 + 語形SubID
書字形ID = 語形ID*256 + 書字形SubID
発音形ID = 語形ID*256 + 発音形SubID

たとえば、語彙素IDが1000の語彙素の子である語形は、語形IDとして32001（1000×32＋1）から32032（1000×32＋32）までの数字を持つことになる。この語形の子である書字形の書字形IDは、8192257（32001×256＋1）から8200448（32032×256＋256）までの数字となる。

したがって、各変化形のIDから親となる見出し語の見出しIDは計算で求めることができる。たとえば、書字形IDが16384257である場合、語形IDは256で割って端数を切り捨てたものである。16384257÷256＝64001.00390625であるから、語形IDは64001となる。また、この語形の語彙素IDは、32で割って端数を切り捨てたものである。64001÷32＝2000.03125であるから、語彙素IDは2000となる。

実際には、ID変換用の関数を用意しているのでデータベース上ではこれを用いて変換することになる。

図 13　見出し語IDの例

親エントリのIDに乗じている数字は、子見出しの最大数を決める定数で、データベースのID変換マスタテーブルに規定されている。この数字は変更される可能性がある。そのため、ID計算に関する全ての処理は、固定した数値を用いず、ID変換マスタテーブル(表 14)の値を使用する。

表 14　ID変換係数マスタテーブル

見出しID	係数
語彙素ID	1
語形ID	32
書字形ID	256
発音形ID	256
(語頭変化形ID)	16
(語末変化形ID)	16
語彙表ID	512

なお、表 14の語頭変化形ID・語末変化形ID・語彙表IDは、後述する語彙表ID生成で利用する数字である。

3.7.3.語彙表ID

活用形・変化形の全てを展開した場合のID（語彙素ID）は、次のように計算される。

$$\text{語彙表ID} = (((\text{書字形ID} * 256 + \text{発音形SubID}) * 16 + \text{語頭変化形subID}) * 16 + \text{語末変化形subID}) * 512 + \text{活用形ID}$$

活用形・変化形の展開が行われるため、語形より下の見出しID（基本形のID）は語彙素IDとは直接に対応しない。式の二重下線部が語頭変化、下線部までが語末変化を反映させたIDに相当する。最後に、活用による変化を反映させるため512を乗じて活用形IDを足している。

図 14に例として形容詞「辛い」の語彙表IDを生成した場合の語彙表IDを図示する。

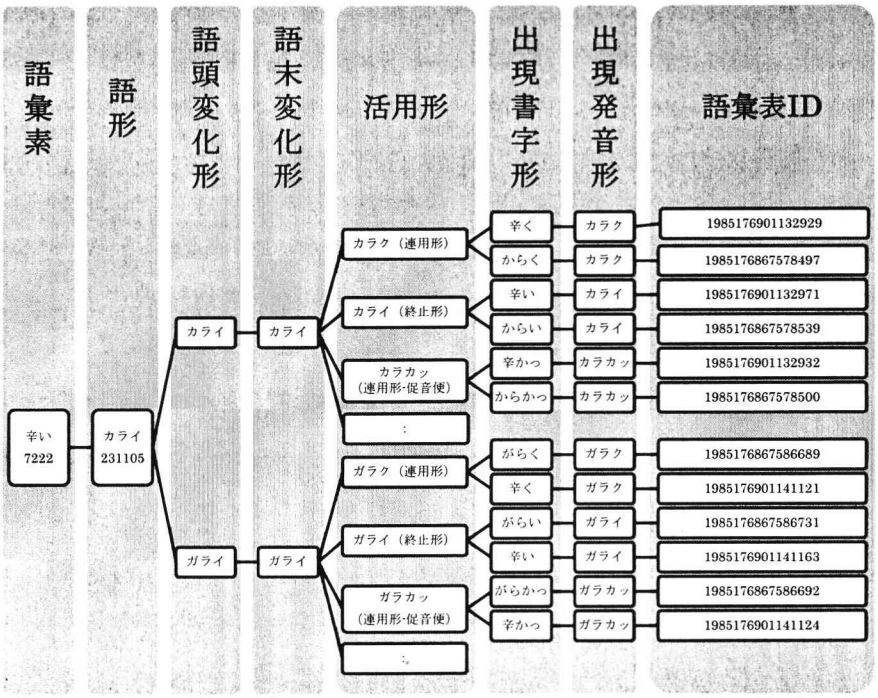


図 14　語彙表 ID 生成の例

3.7.4.見出し表の一意制約

見出し表は、重複した見出しの入力を防ぐために、次の二通りの組み合わせで常にユニークであることを保証する制約が付けられている。これにより重複する見出しは入力することができなくなっている（誤って入力した場合にはロールバックされる）。

この制約は、SQL Server のインデックス付きビュー（Schema Binding）の機能によって実現している。

表 15　見出し表の一意制約

テーブル / 制約	短単位語彙素	短単位語形	短単位書字形	短単位発音形
制約 1	語彙素・語彙素読み・語彙素細分類	語形・品詞・活用型	書字形	
制約 2	語彙素・語彙素読み・語彙素細分類	語形・品詞・活用型	書字形	発音形

なお、単独のテーブル内の見出し制約として、これ以外に短単位語彙素テーブルの次の一意制約がある（3.2.2 参照）。

表 16　語彙素の一意制約

テーブル / 制約	短単位語彙素
語彙素制約	語彙素・語彙素読み・語彙素細分類・類

「類」は「品詞」（語形テーブル）の上位概念であるため、見出し表の一意制約に「類」は含まれていない。

3.8.書字形構成漢字

3.8.1.書字形構成漢字の概要

書字形構成漢字表は、書字形を構成する漢字がどのように読まれているかという情報を持つ。書字形構成漢字表とコーパスを結びつけることにより、コーパス中の漢字の音訓別頻度表を作成することができる。また、単漢字の情報を含む漢字表と結合することにより、常用漢字や教育漢字の音訓がコーパス中の漢字の読みをどれだけ網羅しているかといった情報も得られる。

書字形構成漢字表の実体は辞書データベースの書字形構成漢字テーブルである。書字形構成漢字テーブルは書字形 ID を格納し、短単位書字形テーブルと書字形 ID で対応する。また、書字形 ID 以外に書字形内位置、字種、音訓等種別、音訓を格納している。字種、音訓等種別、音訓については、これら 3 項目の組み合わせで一意となっている漢字テーブルで管理されていて、書字形構成漢字テーブルの字種・音訓等種別・音訓の組み合わせは漢字テーブル内にあるいずれかの字種・音訓等種別・音訓の組み合わせと一致している。

書字形構成漢字テーブル・漢字テーブルの列名等の詳細は資料⑩テーブル一覧を参照のこと。

3.8.2.書字形構成漢字の更新

書字形構成漢字テーブルへのレコードの追加は、トリガを使用した自動処理またはツールを使用した手動処理により行う。

自動処理については、短単位語彙素テーブルと短単位書字形テーブルに作成した自動処理用のトリガにより次の通り実行される。

まず、漢字が含まれる書字形を短単位書字形テーブルに登録すると、書字形構成漢字を生成するトリガが起動し（①）、登録した書字形と仮名形と、関連する短単位語彙素テーブルの情報を元にして（②）、漢字テーブルに登録されたレコードの中から字種・音訓等種別・音訓の組み合わせで最も合致率（精度）の高いものを推測し（③）、その字種・音訓等種別・音訓を書字形構成漢字テーブルに格納する（④）。

また、短単位語彙素テーブルには書字形構成漢字を生成する際に必要な情報（人名・組織名等）が格納されているために、短単位語彙素テーブルのレコードを更新した際にも、短単位語彙素テーブルに関連付けされている短単位書字形テーブルのレコードについて、書字形構成漢字が再生成される。

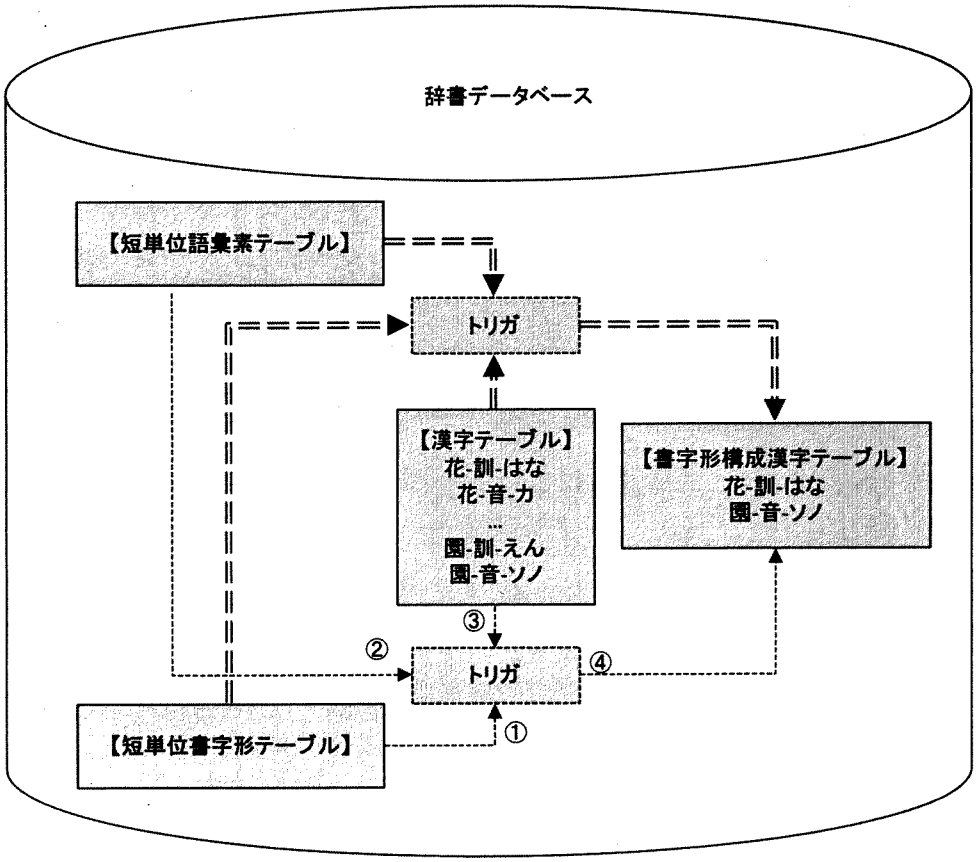


図 15　書字形構成漢字の自動生成概念図

このような自動処理によって生成されたレコードについては、必ず作業者によるチェックが行われ、誤りがあれば修正される。その際に使用されるのが、書字形構成漢字修正ツールである。書字形構成漢字修正ツールについては 5.3 書字形構成漢字修正ツール（46 ページ）を参照。

3.8.3.漢字音訓頻度表生成処理

自動処理によって生成され、手動処理によって整えられた書字形構成漢字テーブルのデータは、漢字音訓頻度表の作成などに利用される。なお、漢字音訓頻度表の生成については専用のエクセルファイルのマクロ処理により行われる。生成条件を与えれば、マクロ処理によって、出現頻度の集計から印刷のために体裁を整える処理まで自動で行われる。

漢字音訓頻度表の生成は次のようなテーブル間の関連性を利用して行われる。漢字テーブルと書字形構成漢字テーブルは字種・音訓等種別・音訓をキーに 1 対多対応している（①）。書字形構成漢字テーブルは書字形 ID を格納しているので、辞書データベースの短単位書字形テーブルと対応している（②）。また短単位テーブルが格納している語彙表 ID からは書字形 ID を算出できるので、短単位テーブルと短単位書字形テーブルは対応している（③）。

以上の関係性により、短単位テーブル内での字種・音訓等種別・音訓の頻度表を容易に生成することができる。

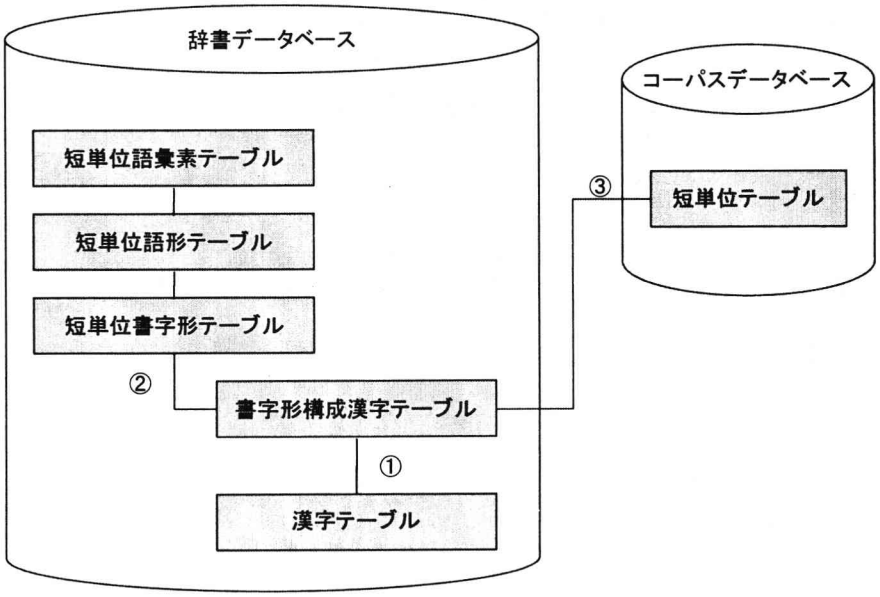


図 16　書字形構成漢字関係のテーブル関連図

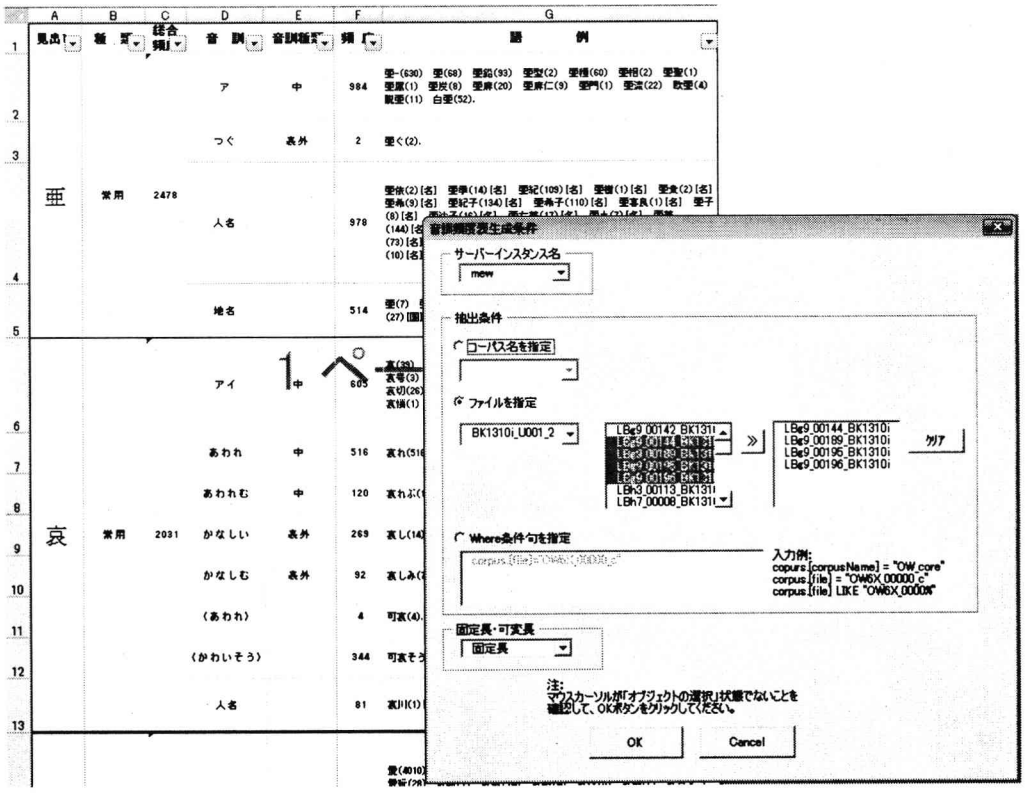


図 17　漢字音訓頻度表生成マクロ

3.9.見出し処理の参考用テーブル

見出し表や語彙表の内容と直接関係するデータではないが、見出し語の入力や修正に当たって作業者が参照する必要のあるデータについても辞書データベース内に格納している。この種のデータには種々のものがあるが、ここでは特に重要な見出し処理の参考用のテーブルについて述べる。

3.9.1.要注意語テーブル

「要注意語」とは、短単位の認定において特に注意を要する語のことで、「要注意語テーブル」はそうした語のリストを格納したものである。要注意語には、付属語扱いする語のリストや、全体で一短単位扱いする例外的な語のリストなどが含まれる。これらについては『形態論情報規程集』にも記載されているほか、辞書データベース用アプリケーションから参照することができるようになっている。

テーブルの仕様については、資料⑩の「要注意語テーブル」を参照。内容については『形態論情報規程集』参照のこと。

3.9.2.要注意誤用例テーブル

「要注意誤用例」は「要注意語」の代表的な用例を登録したテーブルである。一つの要注意語に複数の用例を用意する必要から別テーブルとなっており、ID で関連付けられている。このテーブル内の用例は、要注意語の情報とともに『形態論情報規程集』にも記載されているほか、辞書データベース用アプリケーションから参照することができるようになっている。

テーブルの仕様については、資料⑩の「要注意語用例テーブル」を参照。内容については、『形態論情報規程集』参照のこと。

3.9.3.頻度表

「頻度表」は辞書データベースの見出し語ごとに、コーパスデータベース中の用例数を書き込んだテーブルである。コーパスデータベースの変更を反映するため、ジョブによって定期的に更新されている。

学習用コーパスとして使用されることもある人手修正データについては、個々のコーパスジャンルごとの頻度の内訳が次の例のような書式で記録される。

```
w9:b85:n143:(42832)
```

(コアデータでは白書に 9 例、書籍に 85 例、新聞に 143 例、全コーパスでは 42832 例)

「:」が区切り記号で、アルファベットがジャンルを示す略号、続く数字がジャンル内の用例数、最後の括弧入りの数字がコーパス全体での頻度となっている。コーパスのジャンルを示す略号は、見出し表の「出典」と共通である。

辞書データベース用アプリケーションでも、この形式で各階層の見出し語の品語が表示される。

3.9.4. 語形削除ログ

語形削除ログは、さまざまな理由により語形見出し語を移動したり削除したりした場合に、削除された語形と、削除の日時・ユーザ名などを記録するテーブルである。語形の見出しは、他の見出し表と比べ特に移動が多く外来語の見出し語形などで登録基準を誤りやすいため、特に削除の記録を用意して、削除されたものを再登録することがないように配慮しているものである。

語形削除ログは、見出し表から削除が行われたときにトリガにより自動で記録される(3.3参照)。

3.10. 分類語彙表テーブル

3.10.1. 分類語彙表テーブルの概要

『分類語彙表』とは、国立国語研究所で刊行されている、語を意味によって分類・整理したシソーラス(類義語集)である。UniDicによる形態素解析結果に分類語彙表番号を自動的に付与することを目的に、分類語彙表データベース(『分類語彙表－増補改訂版データベース』)の情報をデータベースに取り込み、UniDicの見出し表と関連付ける(UniDicの見出し語に分類語彙表番号を付与する)作業を行っている。

分類語彙表番号はUniDicの階層では語彙素の階層に付与される。しかし、多義語の場合などに両者の間で一対一の対応をするとは限らない(多対多の関係になる)ため、関連付けのために中間テーブル(分類語彙表関連付けテーブル)を挟んで結合している。

分類語彙表の関連付けには、専用のツールを使用する。分類語彙表ツールについては5.4(48ページ)参照。

3.10.2. 短単位語彙素テーブルとの関連付け

分類語彙表テーブルは中間テーブル(分類語彙表関連付けテーブル)を介して短単位語彙素テーブルと関連付けされている。関連付けには両者の主キーである分類語彙表番号と

語彙素 ID を用いる。表 17・表 18 に分類語彙表関係のテーブルの構成を、図 18 に分類語彙表関係のテーブルと辞書データベース（UniDic）の見出し表との関係を示す。

表 17　分類語彙表テーブル

列名	説明
分類語彙表番号	主キー。分類語彙表データベースの項目と同じ
レコード種別	同上
部門	同上
中項目	同上
分類項目	同上
見出し	同上
見出し読み	同上
更新作業者	（見出し表の共通属性に準ずる）
更新日時	（見出し表の共通属性に準ずる）

表 18　分類語彙表関連付けテーブル

列名	説明
語彙素 ID	短単位語彙素テーブルの ID
分類語彙表番号	分類語彙表の ID
更新作業者	更新作業者名
更新日時	更新日時

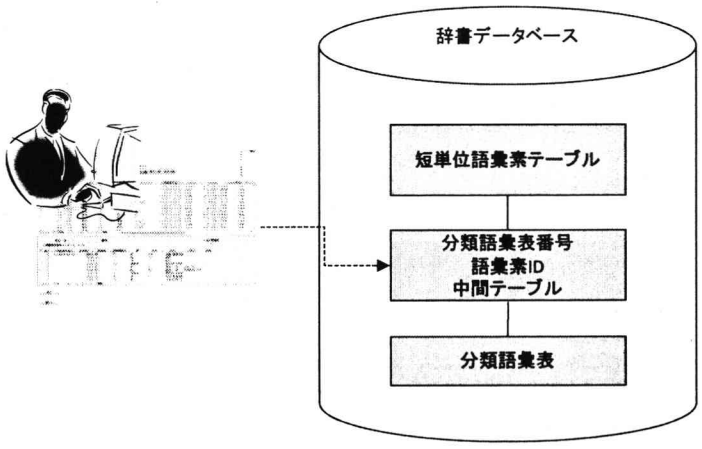


図 18　分類語彙表関係のテーブルと見出し表の関連図

4. コーパスデータベース

4.1.コーパスデータベースの概要

BCCWJ のデータは XML で記述されている。コーパスデータベースでは、この情報を関係データベースの一般的な表で表現するために、「文字表」「短単位表」「文字修正表」「数字タグ表」「ルビ表」「タグ表」の各表に分けて取り込んでいる。形態論情報の処理に直接関連するタグのみ専用テーブルに書き込み、その他のタグは一括してタグ表で保管する。いずれのテーブルもサンプル ID と原文における文字位置をキーとして関連付けられている

コーパスデータベースには各種のコーパスが格納されている。そのうち、人手修正を施したデータをコアデータと呼ぶ。コアデータは形態素解析辞書 UniDic の学習用コーパスとして利用される。コアデータ以外のデータは、見出し表に登録するための未登録語の採集や、コーパスを利用する研究のために用いるデータである。コアデータか否かの区分は短単位テーブルの「コーパス名」によって区別される。BCCWJ のコアデータは「_core」で終わるコーパス名が付けられている。

4.2.コーパスデータベースのテーブル

コーパスデータベース内のテーブルは主に文字テーブルを軸として、サンプル ID と文字開始位置・文字終了位置をキーにして関連付けされている。また、辞書データベースとは語彙表テーブルを介して関連付けされている。これによりコーパスデータベース用アプリケーション・大納言（49 ページ）等のアプリケーションからはコーパスデータベース内のほぼ全てのデータにアクセスできるようになっている。以下にテーブルの一覧とその説明を示す（表 19）。特に重要な短単位テーブルについては 4.3（38 ページ）、長単位・文節テーブルについては 4.4（39 ページ）で詳細を説明する。その他のテーブルについては資料⑩及びサンプルデータ（132 ページ以降）を参照されたい。サンプルデータでは、テキストの同一箇所を例として挙げ、各テーブル上でどのように表現されるかを示している。

表 19　コーパスデータベースのテーブル一覧

テーブル名	説明
文字テーブル	1 レコードにプレーンテキストの 1 文字を格納する、コーパスデータベース内の各テーブルの基準となるテーブル。短単位テーブルや長単位テーブルなどは文字テーブルと常に対応がとれるように更新される。主なフィールドはサンプル ID・文字開始位置・文字終了位置・文字・固定長フラグ・可変長フラグがある。

4.コーパスデータベース

テーブル名	説明
短単位テーブル	1 レコードに 1 短単位、文章（テキスト）を形態素解析した結果を格納するテーブル。主なフィールドにサンプル ID・文字開始位置・文字終了位置・出現書字形・品詞・活用型・語彙表 ID・文開始位置・文終了位置・コーパス名などがある。
数字テーブル	XML における数字タグの情報を格納するテーブル。大納言の対話式数字変換機能を利用して値の修正やレコードの追加・削除が可能である。主なフィールドにサンプル ID・文字開始位置・文字終了位置・数字変換型などがある。
文字修正テーブル	XML における文字修正タグの情報を格納するテーブル。大納言の文字修正機能を利用して値の修正やレコードの追加・削除が可能である。主なフィールドにサンプル ID・文字開始位置・文字終了位置・修正型・原文文字列などがある。
振り仮名テーブル	XML における振り仮名タグの情報を格納するテーブル。大納言の文字修正機能を利用して値の修正やレコードの追加・削除が可能である。主なフィールドにサンプル ID・文字開始位置・文字終了位置・振り仮名などがある。
タグテーブル	XML タグの全ての情報を格納するテーブル。原則としては情報の修正は行われない。主なフィールドにサンプル ID・文字開始位置・文字終了位置・タグ情報がある。
文テーブル	1 レコードに 1 文を格納する、全文検索処理で利用されるテーブル。XML 解析時には存在しないデータである。コーパスデータベースに取り込んだ後、短単位テーブルの文開始位置・文終了位置と対応する形で、データベースのジョブ処理により自動的に生成される。主なフィールドにサンプル ID・コーパス名・文開始位置・文などがある。
語彙表テーブル	1 レコードに 1 短単位を格納する、辞書データベースを利用して生成されるテーブル。未知語等の一部の語を除く短単位テーブルに存在する全ての語を網羅している。辞書データベースの語彙素・語形・書字形・発音形テーブルが更新されると、トリガ処理により語彙表テーブルも更新される。またユニーク ID（語彙表 ID）により、短単位テーブルと対応関係をとる（大納言を使用して対応付けをする）ことによって、辞書データベースの語彙素・語形・書字形・発音形テーブルが更新されると、短単位テーブルも更新される。主なフィールドに語彙表 ID・出現書字形・品詞・活用型などがある。

テーブル名	説明
長単位テーブル	文章（テキスト）を長単位規定に準じて解析した結果を格納するテーブル。1 レコードが 1 長単位になっている。長単位の修正は大納言の長単位モードにより行う。なお、長単位の属性については、長単位語彙表テーブルの中から選択する。主なフィールドにサンプル ID・文字開始位置・文字終了位置・長単位出現書字形・長単位品詞・長単位語彙素などがある。
長単位語彙表テーブル	長単位用の語彙表である。短単位で使われる語彙表テーブルとは異なり、辞書データベースとは連携していない。大納言の長単位モードを利用してデータを整備する。主なフィールドに長単位出現書字形・長単位品詞・長単位活用型などがある。
文節テーブル	文節情報が格納されるテーブル。大納言の長単位モードを利用して修正を行う。主なフィールドにサンプル ID・文字開始位置・文字終了位置・文節などがある。

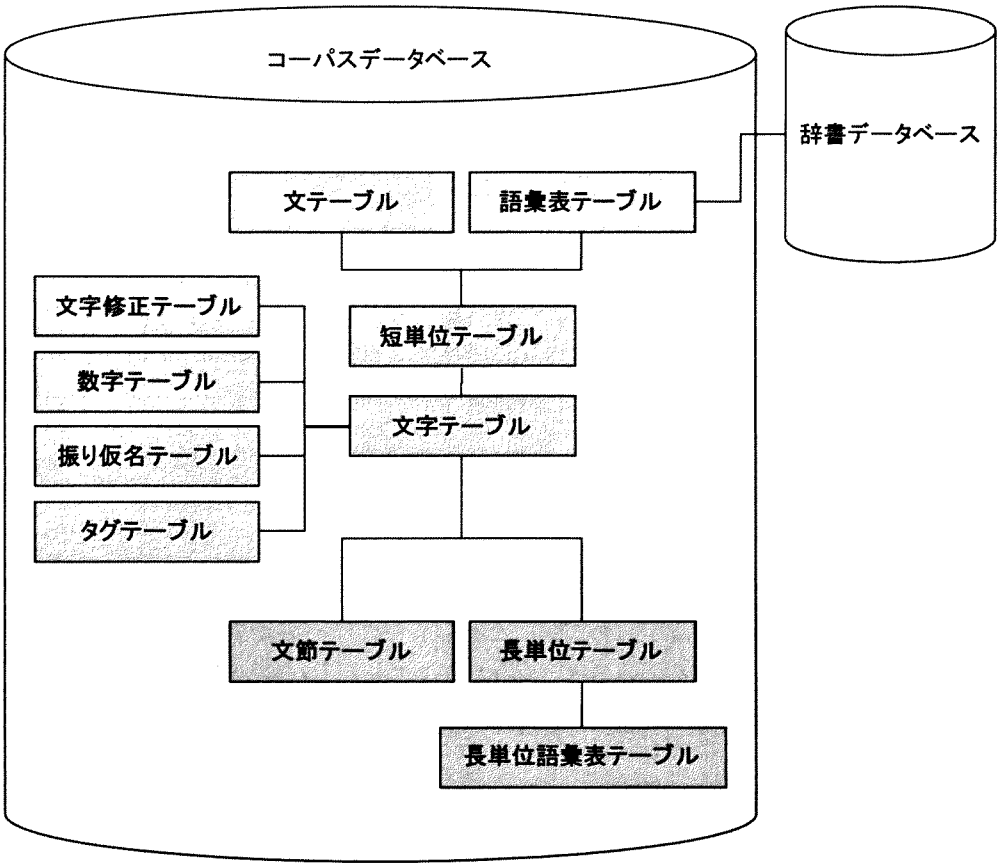


図 19　コーパスデータベースのテーブル関連図

4.3.短単位テーブル

短単位テーブルは形態素解析結果を取り込んだもので、コーパスデータベース内でも最も重要な役割をもつテーブルであり、SQL 文からで直接利用することも多い。利用に際して必要となる情報を表 20 に示す。「（冗長）」とした項目は、データ利用の便宜上、短単位テーブル内に保持しているものの、他のマスタテーブルから取得可能な情報である。

表 20　短単位テーブルの列名

項目	説明	形態素解析の出力	取り込み時に必須	区分
コーパス名	コーパス名（ジャンル別等）		○	基本となる出典情報
サンプル ID	BCCWJ のサンプル ID	○	○	
連番	サンプル内の並び順	○※	○	
文境界	文頭（B）またはそれ以外(I)	○	○	
文字開始位置	文字テーブルの開始 ID		○	文字表・その他のテーブルとの接続用
文字終了位置	文字テーブルの終了 ID		○	
語彙素読み	当該短単位の語彙素読み	○	○	基本となる形態素情報（基本 8 属性）
語彙素	当該短単位の語彙素	○	○	
語彙素細分類	当該短単位の語彙素細分類	○	○	
品詞	当該短単位の品詞	○	○	
活用型	当該短単位の活用型（簡略活用型）	○	○	
活用形	当該短単位の活用形（簡略活用形）	○	○	
出現書字形	語形変化・活用後の書字形	○	○	
出現発音形	語形変化・活用後の発音形	○	○	
語彙表 ID	展開した語彙表の ID（展開後の語として一意）			基本となる形態素 ID
語彙素 ID	対応する短単位語彙素の ID			コーパス利用のための追加形態素情報（冗長）
語種	当該短単位の語種	○		
語形	語形（語形変化・活用前の基本形）	○		
文開始位置	文テーブルの開始 ID			文テーブルとの接続用
文終了位置	文テーブルの終了 ID			
固定長フラグ	BCCWJ の固定長サンプル内か否か			コーパス利用のための追加出典情報（冗長）
可変長フラグ	BCCWJ の可変長サンプル内か否か			
学習フラグ	学習用コーパスとしての採否情報			学習用コーパスとしての情報
UpdUser	最終更新ユーザ名			更新情報
UpdDate	最終更新日時			

※出力の並び順

4.4.長単位テーブルと文節テーブル

長単位は、BCCWJ の形態論情報として付与される言語単位の一つで、文節をもとに、そこから付属語等を切り出したものに相当する。一つの長単位は、一つの短単位または複数個の短単位の連続となる。BCCWJ における長単位・文節の定義については『形態論情報規程集』を参照のこと。

短単位と長単位・文節は、表 21 のような関係にあり、文節境界は常に長単位境界であり、文節・長単位境界は常に短単位境界となる。また、文節や長単位は短単位の連続からなる。ただし、注釈的な括弧などにより、長単位が短単位の連続とならない場合がある。短単位と長単位は、語彙素・品詞・活用型等の情報をもつが、文節は境界のみを記録している。

表 21　短単位・文節境界・長単位の例

短単位境界	短単位	文節境界	長単位境界	長単位
B	文化	B	B	文化庁文化交流使事業
B	庁			
B	文化			
B	交流			
B	使			
B	事業			
B	は		B	は
B	,		B	,
B	芸術	B	B	芸術家
B	家			
B	,		B	,
B	文化	B	B	文化人等
B	人			
B	等			
B	,		B	,
B	文化	B	B	文化
B	に			に
B	携わる	B	B	携わる
B	人々	B	B	人々
B	に			に
B	,		B	,
B	一定	B	B	一定期間
B	期間			

長単位はコーパスに出現したものを単位として認めるという形を取っており、コーパスから切り離した見出し表としては管理しない。そのため形態論情報データベースではコーパスデータベースの中でのみ取り扱われ、辞書データベースとは直接関係しない。

長単位に関係するテーブルとしては、長単位テーブルと文節テーブル、長単位語彙表テーブルがある。

長単位テーブルは、出現した長単位の情報を格納するテーブルであり、語彙素・品詞・活用型などの情報が、短単位の情報を利用して付与される（資料⑩、サンプルデータ⑭（135ページ）参照）。文節テーブルは文節境界を記録するためのテーブルである（資料⑩、サンプルデータ⑮（136ページ）参照）。長単位テーブルと文節テーブルは、短単位テーブルと同様に、文字表テーブルのサンプルIDと文字開始位置・文字終了位置によって関連付けされている。なお、長単位テーブルと文節テーブルは先に述べたような密接な関係にあるが、現在のデータベース上では別々のテーブルとして管理している。

長単位語彙表テーブルは、一度出現した長単位を記録して、長単位付与作業に利用するためのテーブルである（6.8.2（83ページ）、及び、資料⑩、サンプルデータ⑯（136ページ）参照）。長単位テーブルと長単位語彙表テーブルは属性（長単位出現書字形・長単位品詞等）で関連付けされている。

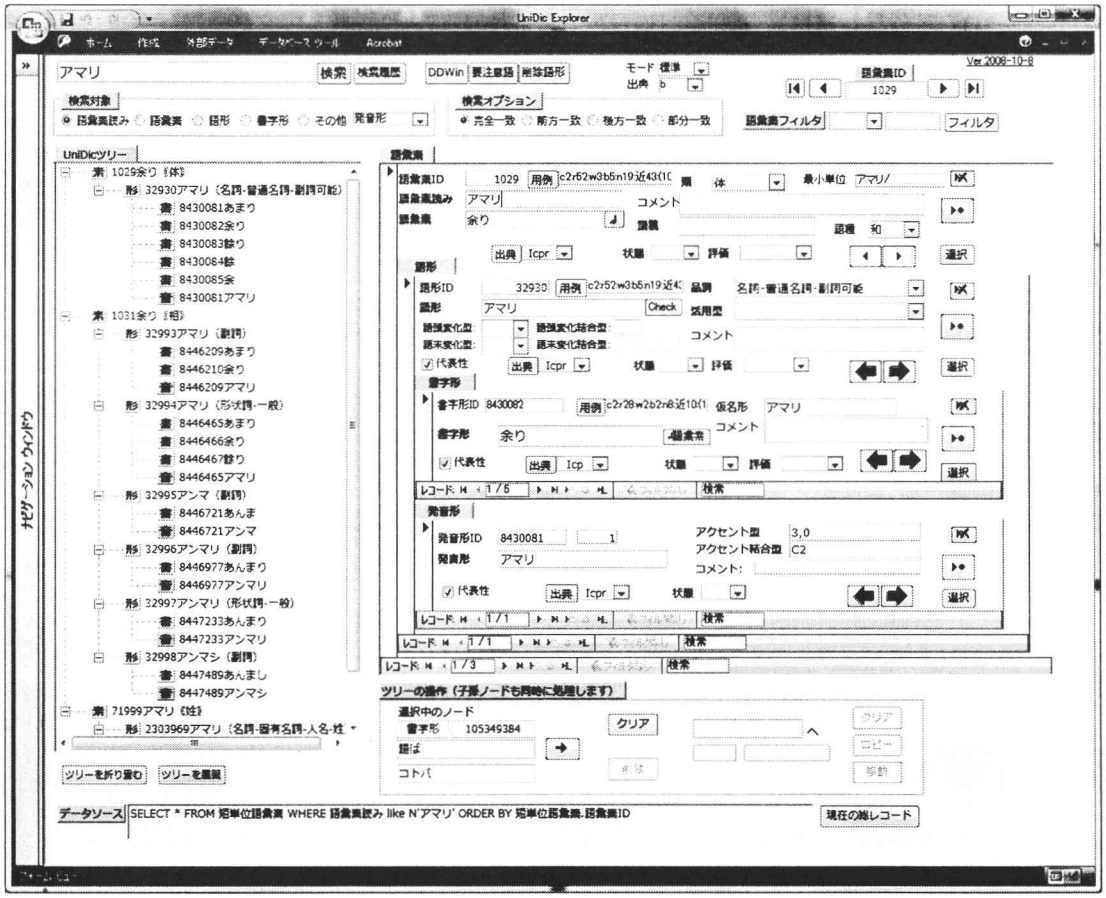
5. 辞書データベース用アプリケーション

5.1. 概要

辞書データベースへの登録・修正を行うアプリケーションとして、辞書管理ツール「UniDic Explorer」がある。また、特定目的のツールとして「書字形構成漢字修正ツール」「分類語彙表ツール」がある。この 3 種のアプリケーションについてその機能、処理内容を説明する。

5.2. 辞書管理ツール UniDic Explorer

辞書管理ツール「UniDic Explorer」は辞書データベースに見出し語を追加するための中心となるツールである。



見出し語の追加・修正作業時には、見出し語表の階層をそのまま表示し、修正が可能となっている。以下、その機能について説明する。

5.2.1.見出し語の検索

UniDic Explorer では、各階層の見出し語や関連する情報をもとに、見出し表に登録された語を検索・表示することができる。

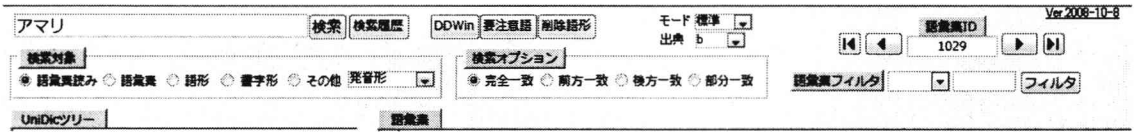


図 20　UniDic Explorer の検索用コントロール

検索対象としては、使用頻度の高い「語彙素読み」「語彙素」「語形」「書字形」のほか、「その他」を選択して発音形や見出しに付けられたコメントなどを検索することができる。この際、検索オプションとして条件を「完全一致」「前方一致」「後方一致」「部分一致」から検索できる。語彙素 ID を入力することで、直接語彙素を指定することも可能である。

左ペインには検索した語が UniDic の階層を反映したツリー構造で表示され、右ペインには各階層の見出し語が、階層構造をそのまま反映した形で表示される。

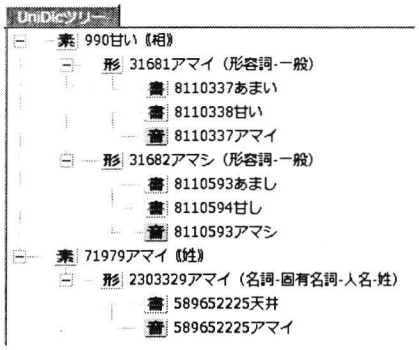


図 21　UniDic の階層を反映したツリー

ツリーには階層を示すアイコンと各見出し語の ID、各階層の代表的な項目が表示される。項目は、語彙素見出しでは語彙素と類、語形見出しでは語形と品詞、書字形見出しでは書字形、発音形見出しでは発音形である。ツリーの項目をクリックすると、当該レコードが選択され、右ペインに表示されて編集が可能になる。

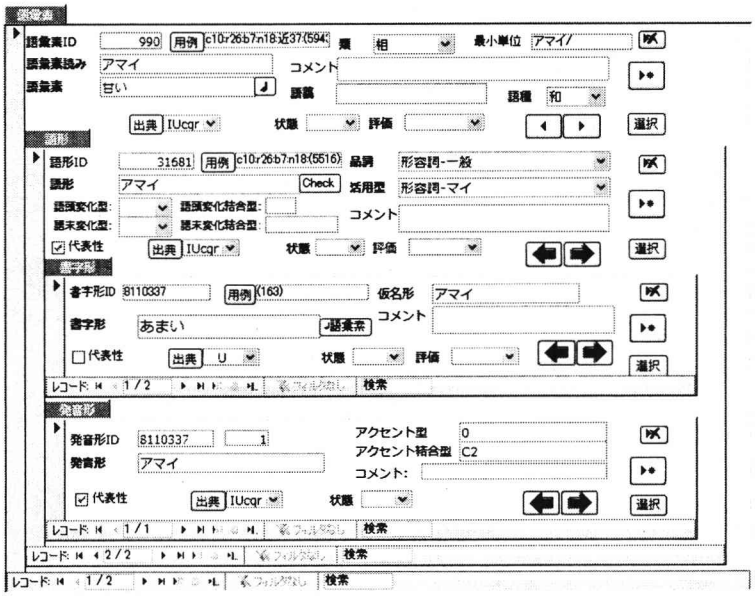


図 22　UniDic の階層を反映したレコード表示

5.2.2.見出し語の追加

見出し語の追加は、各見出し階層画面の ▸▸ ボタンによって行う。このボタン押下時に、ID は所定の手続きにより自動で計算され入力される（3.7.2 参照）、こののち新規見出し語の入力が可能になる。

見出し表の制約により、見出し語は必ず親となる見出し語から追加する必要がある。また、見出し語を削除する場合には、その見出し語の子となっている見出し語を全て削除しなければならない。

なお、画面上部の「出典」を選択しておくことで、新規レコードの出典が自動的に入力される。出典の選択肢は出典テーブルと関連付けられている。また、画面上部の「モード」で「仮登録」を選択すると、新規レコードの状態として「仮」が自動的に入力される。

5.2.3.見出し語の修正

データベースのレコードを表示するコントロールは、そのままデータベース上の項目と関連付けられているため、画面上での修正した結果はそのままデータベースレコードの修正として反映される。アップデート処理は、修正したレコードから他に移動したときに行われる。

なお、画面上部の「モード」で「閲覧」を選択すると、誤って修正することを禁止する閲覧モードとなり、レコードの修正ができなくなる。

5.2.4.見出し語の移動・コピー

ツリーの項目を選択するか、右ペインの「選択」ボタンを押下することにより、項目が選択され、画面下のツリー操作用コントロールに選択項目が表示される（図 23 の①）。この状態で「→」ボタンを押下すると、右側のコントロールが利用可能になり、当該項目のコピー・移動を行うモードとなる（②）。もう一度ツリーの項目を選択するか、右ペインの「選択」ボタンを押下することにより、移動・コピー先が右側に指定される（③）。その後、「コピー」ボタンを押下すると当該項目をコピー、「移動」を押下すると当該項目を移動する。

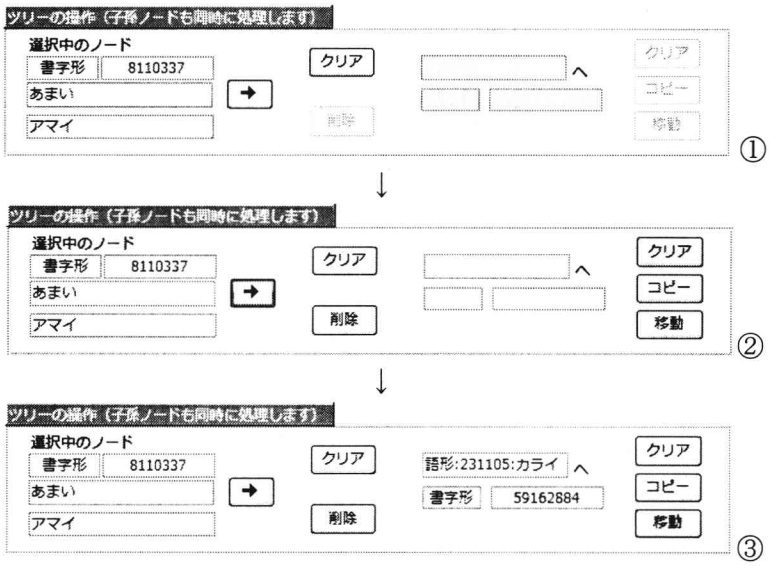


図 23　見出し語の移動・コピー

移動・コピーは当該見出し語だけでなく、子や孫となる見出し語全体をまとめて行われる。なお、②の状態で「削除」ボタンを押すことにより、当該の見出し語を子や孫となる見出し語ごと全て削除することもできる。

同一見出しの元にコピーする場合、一意制約に対応するため、同一見出し語の場合には主となる見出しの後に「（コピー）」の文字を付与したものがコピーされる。

5.2.5.参考情報の参照

「要注意語」などの見出し処理の参考用テーブルは、UniDic Explorer の画面上から呼び出して閲覧することができる（3.9 見出し処理の参考用テーブル・32 ページ参照）。

検索用テキストボックスに検索語を入力語、画面上部の「要注意語」「削除語形」等のボタンを押下することにより、該当する語の情報を表示することができる。

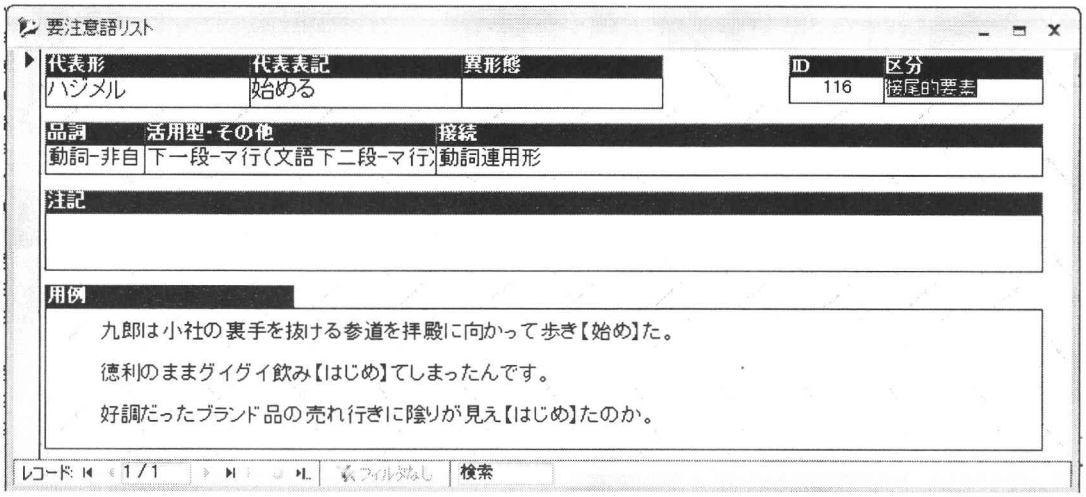


図 24　要注意語テーブルの参照

頻度表の情報（コーパス中の頻度）は右ペインの各階層の見出し語の部分に常に表示されている。頻度情報の横の「用例」ボタンを押下することで、当該語のコーパス中の用例を文脈付きで全て表示することができる。

図 25　頻度表の情報と用例参照ボタン（書字形）

図 26　コーパス中の用例の参照

5.3. 書字形構成漢字修正ツール

自動生成処理（3.8 書字形構成漢字・29 ページ参照）によって書字形構成漢字テーブルに追加されたレコードは、書字形構成漢字修正ツールを使用してチェックする。データが誤っている場合には、正しい情報に修正する。必要であれば、漢字テーブルへのレコードの追加も行う。

書字形構成漢字テーブルは、漢字についての情報（字種・音訓種別・音訓）以外に、自動処理時の精度情報と、手動処理の際に入力する確定フラグを格納している。精度情報については、自動処理によって書字形構成漢字のレコードが生成された際の、結果の確からしさを数値で表している（最低 0~最高 1）。また、確定フラグは作業者によるチェックや修正作業が終了したことを表している。

書字形構成漢字テーブル内で確定フラグが立っていないレコードについては、夜間のジョブによって再生成処理が行われる。作業者によって漢字テーブルに新しくレコードが追加されれば、再生成処理によってこれまで誤っていたものに正しい漢字の情報が付与される可能性があるためである。

図 27　書字形構成漢字修正ツール

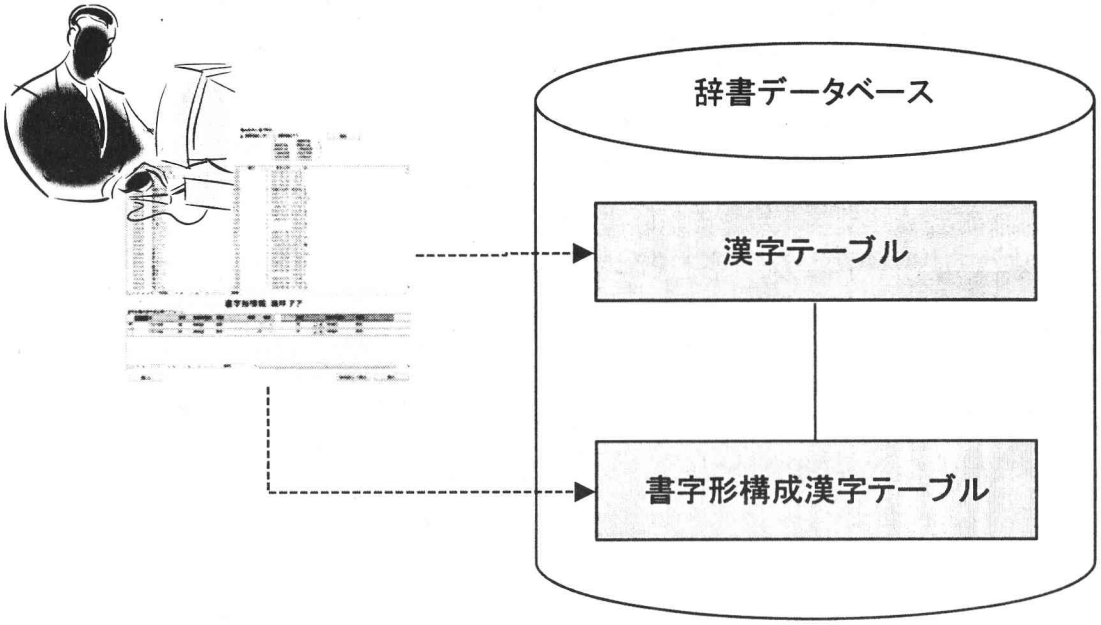


図 28　書字形構成漢字修正ツールの概念図

5.4. 分類語彙表ツール

分類語彙表テーブルと語彙素テーブルの関連付け作業（3.10 分類語彙表テーブル・33 ページ参照）には、分類語彙表ツールを使用する。分類語彙表ツールを使用して、関連付けテーブルへのレコードの追加や削除などを行う。

分類語彙表ツールにおける分類語彙表の検索項目は、分類番号、見出し、見出し読み、分類語彙表番号があり、それぞれ完全一致、前方一致、後方一致による検索を行うことができる（①）。

分類語彙表の検索結果は②に表示される。また②で選択した分類語彙表テーブルのレコードと対応している、または対応付けの候補として考えられる短単位語彙素が③に表示される。なお、語彙素読みまたは語形が見出し読みと一致するものを候補としている。

作業者は関連付けする語彙素を③で選択し、実行ボタンを押す（④）。すると分類語彙表関連付けテーブルにレコードが追加され、短単位語彙素テーブルと分類語彙表番号とが関連付けされる。

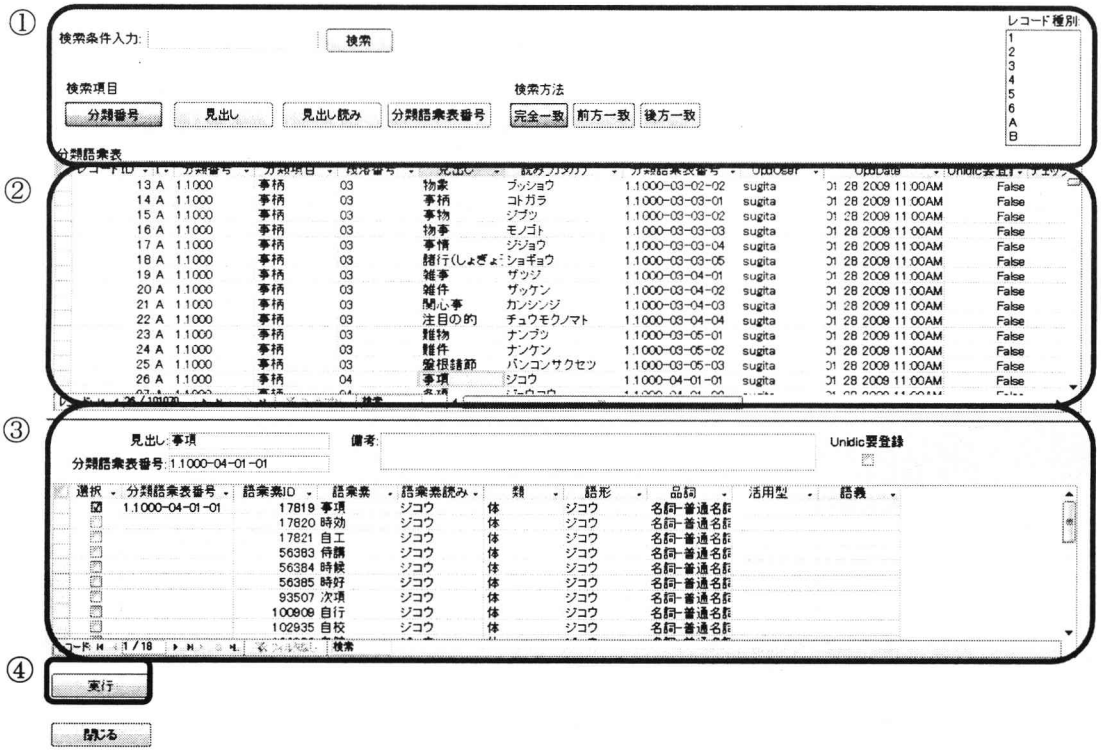


図 29　分類語彙表ツール

6. コーパスデータベース用アプリケーション・大納言

6.1.大納言の概要

大納言は 1 億語規模の短単位とそれに付随するデータを格納するコーパスデータベース内の各テーブルに対する検索、更新を行うためのツールである。

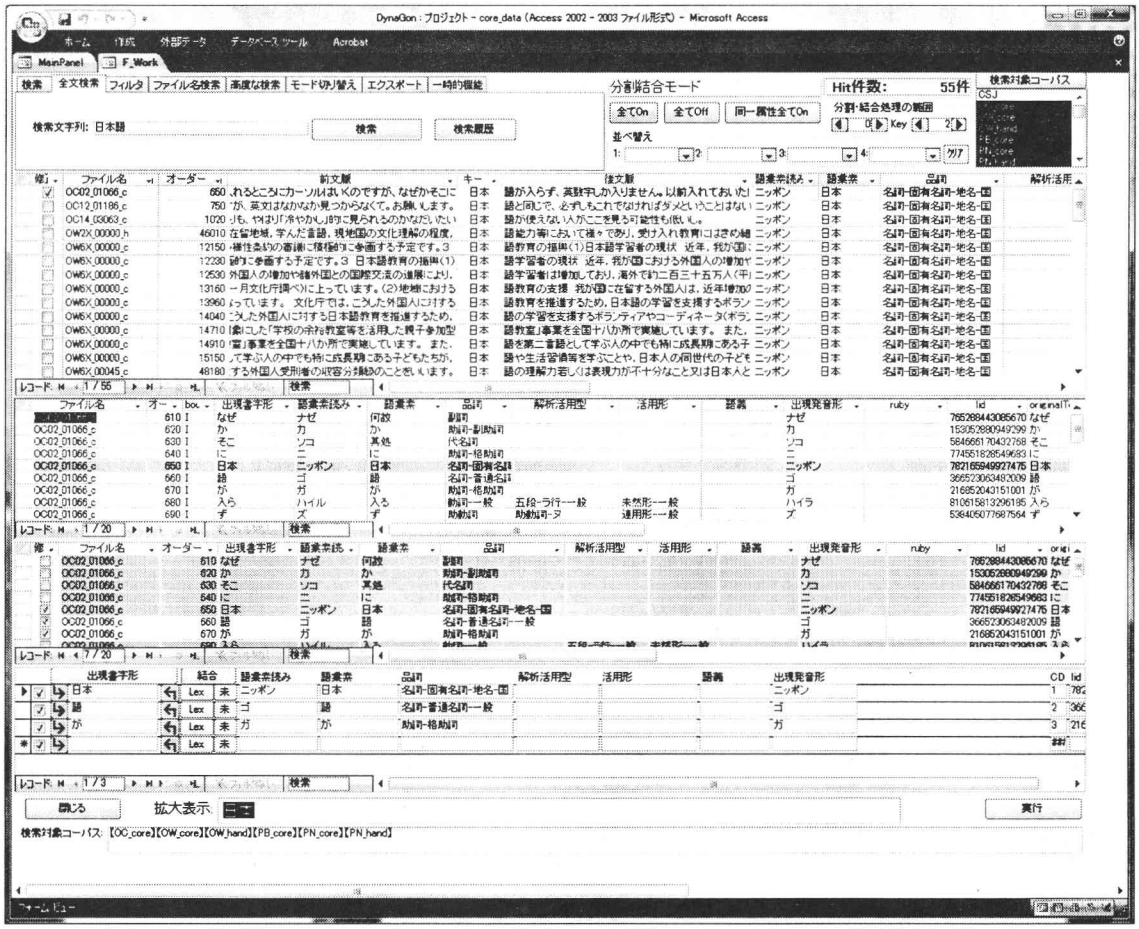


図 30　大納言の基本操作画面

大納言は、表面上は Access アプリケーションだが、実体はデータベースに格納されたストアドプロシージャや関数等と連動した一連のシステムとなっている。大納言で使用されている主なソフトウェアは以下の通りである。

OS	Windows Server 2003 R2
データベース	SQL-Server2005
クライアントソフトウェア	Microsoft Access 2000 以上

6.2.メイン作業画面

大納言のメイン作業画面を以下に示す。

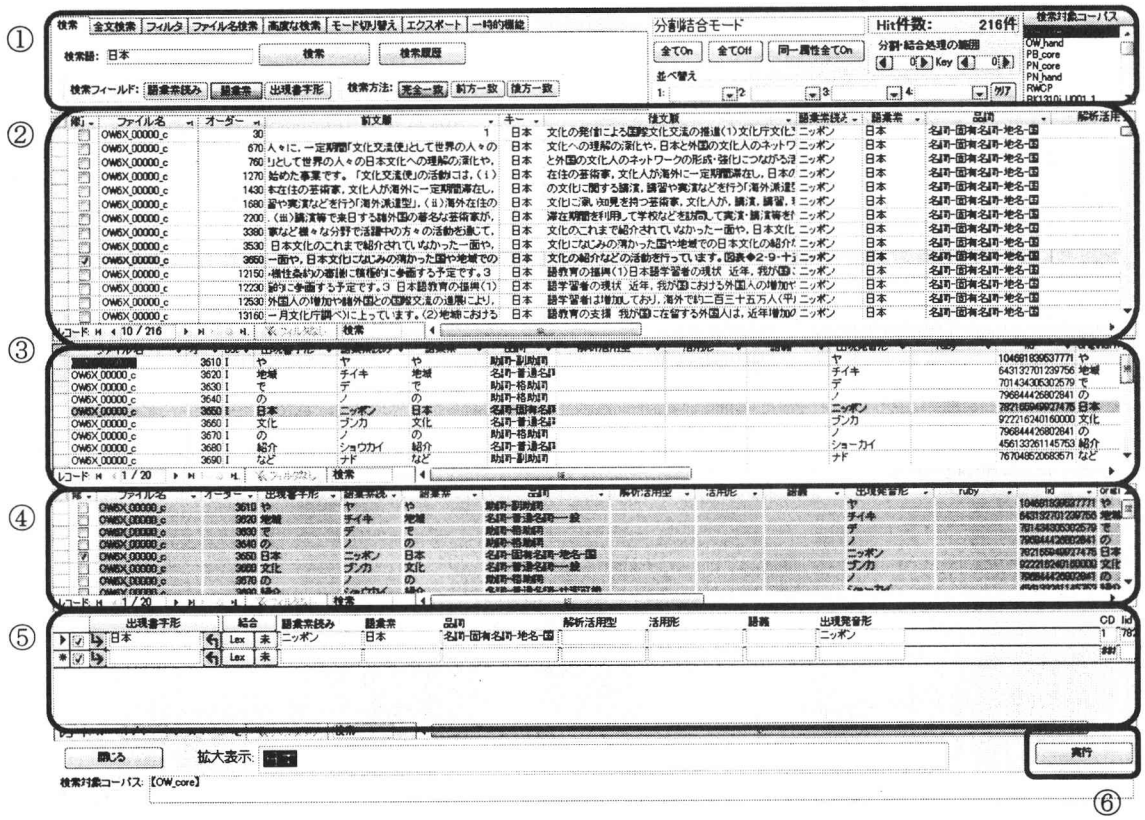


図 31　大納言のメイン操作画面

①コントロール部

検索条件の入力やソート項目の指定、モード切り替え等の基本的な操作を行う部分。

②KWIC 表示部

検索結果が表示される。分割結合や対話式数字変換処理等の処理する語の選択はここで行う。

③周辺語情報表示部

KWIC 表示部（②）で選択中の語の前後（周辺）の語の情報が表示される。また、KWIC 表示部では表示していない数字情報や文字修正情報、振り仮名情報等も表示される。

④処理範囲指定部

KWIC 表示部（②）と組み合わせて使用する。KWIC 表示部（②）で選択した語の処理範囲を指定する。

⑤修正内容指定部

正しい語の区切り位置を指定する。また、語の属性情報を語彙表から選択する形で入力する。分割結合等の処理をした場合は、②で選択された語について④の範囲が⑤に置き換わる。

⑥実行ボタン

実行ボタンを押すことでストアドプロシージャが起動し、コーパスデータベース内のテーブルの値が更新される。更新前・更新中・更新後には文脈チェックを行い、データが不正に書き変わらないかをチェックしている。問題があった場合、処理はロールバックされる。

6.3.大納言の機能

大納言の主な機能としては、以下のものがある。

6.3.1.検索機能

大納言では以下の検索方法によりデータベース内を検索することができる。検索結果はKWICが付与された状態で表示される。

・短単位検索

- 語彙素読み（完全一致・前方一致・後方一致）の検索
- 語彙素（完全一致・前方一致・後方一致）の検索
- 出現書字形（完全一致・前方一致・後方一致）の検索

・全文検索

短単位の境界を意識することなく、出現書字形を検索することができる。検索条件に正規表現を使用することもできる。検索には全文検索用の文テーブルを使用する。全文検索システムのロジックは後述する。

・サンプルID検索

サンプルIDを指定して検索する。複数のサンプルIDを指定することもできる。

・高度な検索

3語の繋がりまででであれば、検索条件を自由に指定して検索することができる。理論上はコーパスデータベース、辞書データベースに保存されているあらゆるデータに対して検索することが可能である。また、検索条件は保存することができ、作業者間で検索条件を共有することができる。この仕組みによって、管理者が作成した複雑な検索条件を作業者も簡単に利用することができる。

6.3.2.ソート機能

検索結果のKWICを並び替えて表示することができる。ソート項目は最大4つまで指定することができる。

6.3.3.同一属性一括処理機能

同じ属性を持つ語については、一括で更新処理を行うことができる。この処理については 6.5.3 同一属性レコードの一括処理（66 ページ）を参照。

6.3.4.文字修正機能

文字テーブルのデータを修正することができる。データ修正時には関連するテーブルのデータも修正され、整合性が維持される。文字修正機能を利用する際は大納言を文字修正モードに切り替えて行う。

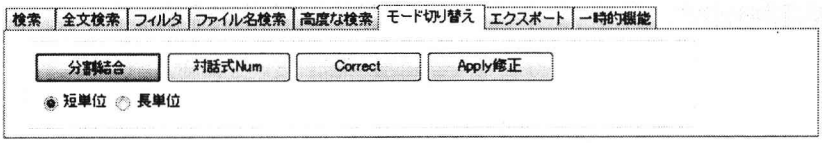


図 32　大納言のモード切替ボタン

6.3.5.対話式数字変換機能

手作業による数字変換処理をサポートする。データ修正時には関連するテーブルのデータも修正され、整合性が維持される。対話式数字変換機能を利用する際は大納言を対話式数字変換モードに切り替えて行う。内容については 6.6（76 ページ）を参照。

6.3.6.長単位分割結合機能

長単位の境界と属性を修正することができる。長単位の属性は長単位語彙表テーブルにあるものから選択する。短単位語彙表とは異なり、長単位語彙表テーブルは辞書データベースとは連携しておらず、コーパスデータベースのみで管理する。辞書データベースの更新は、長単位語彙表テーブルに影響しない。長単位分割結合を利用する際は大納言を長単位分割結合モードに切り替えて行う。

6.3.7.データのインポート機能

形態素解析によって出力された解析結果のテキストと関連するデータを、データベース上のテーブルにインポートすることができる。取り込みできるデータは短単位データ（テーブル）・文字データ（テーブル）・文字修正データ（テーブル）・タグ（テーブル）・数字データ（テーブル）・振り仮名データ（テーブル）である。

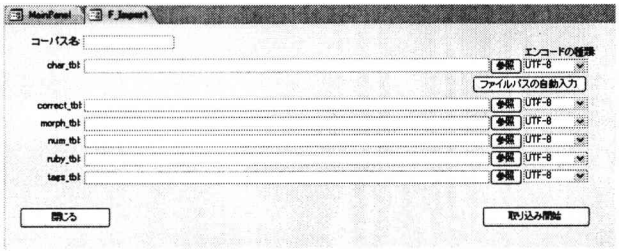


図 33　大納言のデータインポート機能

文字修正データ・振り仮名データ・数字データ・タグデータは必ずしもインポートする必要はない。タグデータは、大納言を使用した人手修正後にデータベース内のデータを使用して XML 文書を再構成してエクスポートする場合にのみ必要となる。

なお、大量のデータを一度にインポートする必要がある場合には、DBMS の管理ツールによって手動で読み込む必要がある。

6.3.8.データの削除機能

コーパスデータベースは複数のテーブルが連動しているので、データの削除を適切に行わないとテーブル間の連動性が失われてしまう危険があるが、大納言ではデータの削除を安全に行うことができる。

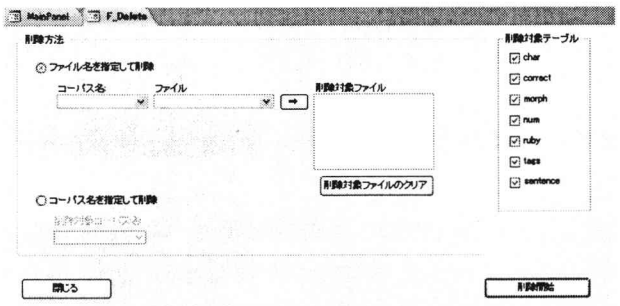


図 34　大納言のデータ削除機能

6.3.9.エクスポート機能

短単位検索・全文検索・サンプル ID 検索・高度な検索での検索結果の KWIC をテキスト形式（符号化方式は UTF-16LE）で保存することができる。

6.3.10. 処理時の文脈チェック機能

同時実行性を低下させないために、テーブルのロックは最小限にしている。そのため、複数の作業者が同時に更新処理した場合でもオリジナルの文が失われることがないよう、処理の過程で文脈のチェックが頻繁に行われる。

6.3.11. 文節修正機能

文節を修正することができる。文節修正機能を利用する際は大納言を長単位分割結合モードに切り替えて行う。

6.3.12. データの保護

大納言は作業者が複数いることを想定して、各作業者専用の作業テーブル（一時テーブル）を使用して作業内容を管理している。大納言を使用した操作内容は作業テーブルに反映され、短単位テーブル等の更新はデータベースに登録されたストアドプロシージャが作業テーブルのデータを利用して行う。作業テーブル以外の短単位テーブル等のテーブルはユーザから隔離されているので、作業者の誤入力や誤操作などのトラブルからデータが守られるようになっている。また、一連のデータ更新処理はトランザクション処理で行われるので、処理の過程でトラブルが起こった場合でもデータの対応関係が維持される。

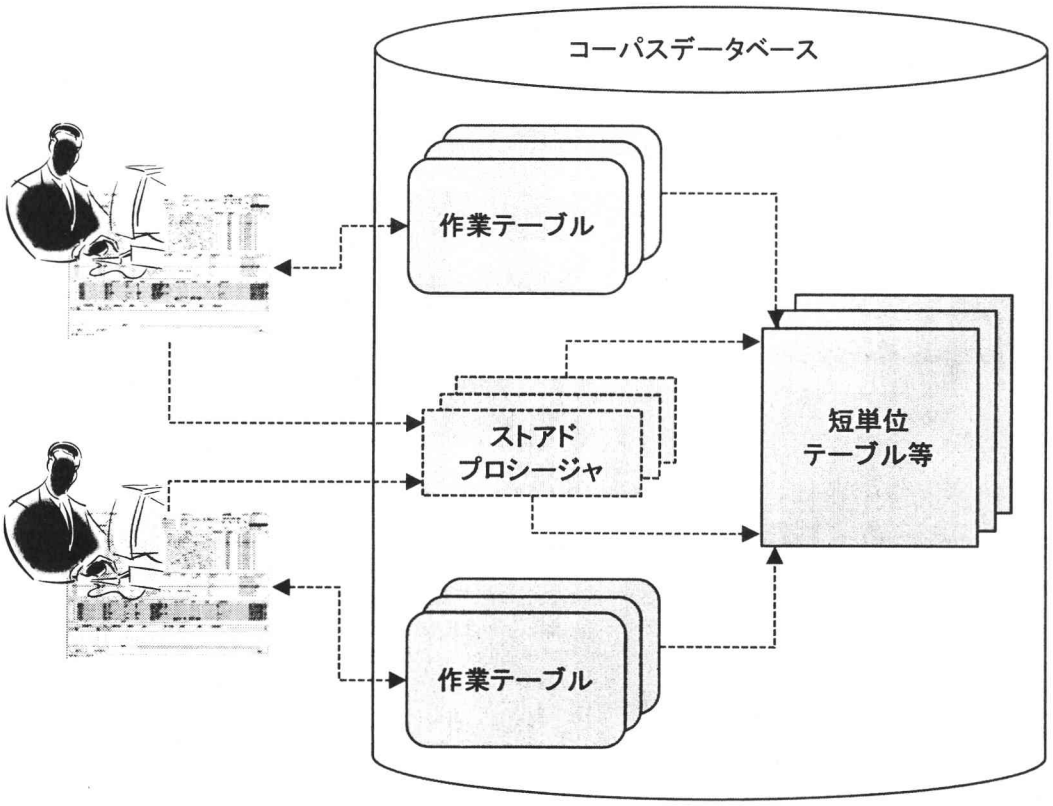


図 35　作業テーブルを使用したデータの隔離

6.4. 検索機能

6.4.1. 検索処理の概要

コントロール部のうち、検索に使用される部分について説明する。検索方法は大きく分けて 4 種類ある。

- 短単位検索
- サンプル ID 検索
- 全文検索
- 高度な検索

短単位検索は短単位テーブルの語彙素、語彙素読み、書字形に対する検索を行う。また、それぞれ検索方法として前方一致・後方一致・完全一致を選択することができる。

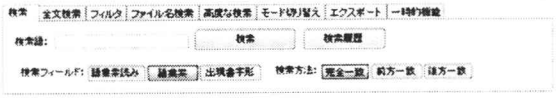


図 36　大納言の検索用コントロール

短単位検索結果の表示例を以下に示す。短単位検索は修正すべき短単位があらかじめわかっている場合や同一属性一括処理をする場合などに有効である。

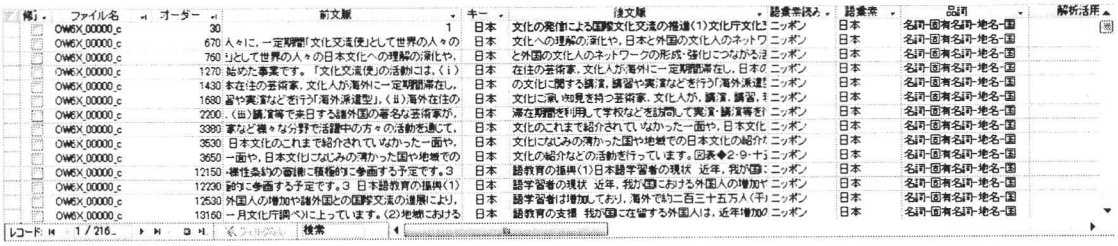


図 37　短単位検索による検索結果の例

サンプル ID 検索は、短単位テーブルのサンプル ID について検索を行う。検索対象のサンプル ID を複数指定することもできる。

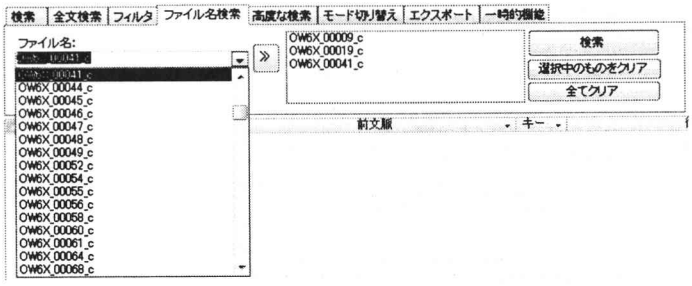
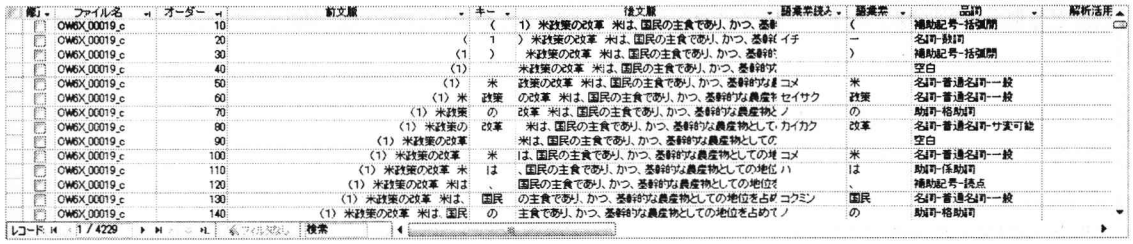


図 38　サンプル ID 検索の例

サンプル ID 検索結果の表示例を以下に示す。サンプル ID 検索は、特定のサンプルについて先頭から順番に短単位をチェックしていく場合などに有効である。



順	ファイル名	オーダー	前文脈	キー	後文脈	語彙素読み	語彙素	品詞	解析活用
	OW6X_00019_c	10		（	1）米政策の改革　米は、国民の主食であり、かつ、基幹	（	（	補助記号-括弧開	
	OW6X_00019_c	20		1	）米政策の改革　米は、国民の主食であり、かつ、基幹	イチ	一	名詞-数詞	
	OW6X_00019_c	30	（1	）	米政策の改革　米は、国民の主食であり、かつ、基幹的		）	補助記号-括弧閉	
	OW6X_00019_c	40	（1）		米政策の改革　米は、国民の主食であり、かつ、基幹的な			空白	
	OW6X_00019_c	50	（1）	米	政策の改革　米は、国民の主食であり、かつ、基幹的な農	コメ	米	名詞-普通名詞-一般	
	OW6X_00019_c	60	（1）米	政策	の改革　米は、国民の主食であり、かつ、基幹的な農産物	セイサク	政策	名詞-普通名詞-一般	
	OW6X_00019_c	70	（1）米政策	の	改革　米は、国民の主食であり、かつ、基幹的な農産物と	ノ	の	助詞-格助詞	
	OW6X_00019_c	80	（1）米政策の	改革	米は、国民の主食であり、かつ、基幹的な農産物として	カイカク	改革	名詞-普通名詞-サ変可能	
	OW6X_00019_c	90	（1）米政策の改革		米は、国民の主食であり、かつ、基幹的な農産物としての			空白	
	OW6X_00019_c	100	（1）米政策の改革	米	は、国民の主食であり、かつ、基幹的な農産物としての地	コメ	米	名詞-普通名詞-一般	
	OW6X_00019_c	110	（1）米政策の改革 米	は	、国民の主食であり、かつ、基幹的な農産物としての地位	ハ	は	助詞-係助詞	
	OW6X_00019_c	120	（1）米政策の改革 米は	、	国民の主食であり、かつ、基幹的な農産物としての地位を	、	、	補助記号-読点	
	OW6X_00019_c	130	（1）米政策の改革 米は、	国民	の主食であり、かつ、基幹的な農産物としての地位を占め	コクミン	国民	名詞-普通名詞-一般	
	OW6X_00019_c	140	（1）米政策の改革 米は、国民	の	主食であり、かつ、基幹的な農産物としての地位を占めて	ノ	の	助詞-格助詞	

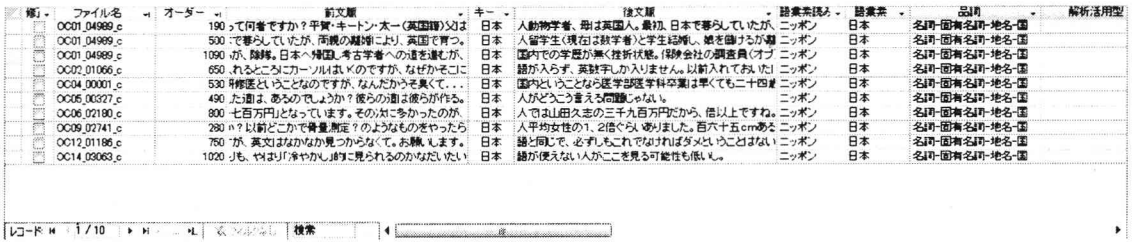
レコード 1 / 4229　検索

図 39　サンプル ID 検索による検索結果の例

全文検索については、文テーブルを使用して検索を行う（処理の詳細については後述）。検索文字列に正規表現を使用することもできる。

図 40　全文検索の例（正規表現）

全文検索検索結果の表示例を以下に示す。全文検索は、誤解析などで短単位がどこで区切られているかわからない場合や、正規表現を利用したパターンマッチングを行いたい場合などに有効である。



順	ファイル名	オーダー	前文脈	キー	後文脈	語彙素読み	語彙素	品詞	解析活用型
	OC01_04989_c	190	って何者ですか？平賀・キートン・太一（英国籍）父は	日本	人動物学者、母は英国人。最初、日本で暮らしていたが、	ニッポン	日本	名詞-固有名詞-地名-国	
	OC01_04989_c	500	で暮らしていたが、両親の離婚により、英国で育つ。	日本	人留学生（現在は鉄学者）と学生結婚し、娘を儲けるが離	ニッポン	日本	名詞-固有名詞-地名-国	
	OC01_04989_c	1090	が、陳峰。日本へ帰国し考古学者への道を進むが、	日本	国内での学歴が無く採折状態。保険会社の調査員（オブ	ニッポン	日本	名詞-固有名詞-地名-国	
	OC02_01066_c	650	れるところにカーソルはいくのですが、なぜかそこに	日本	語が入らず、英数字しか入りません。以前入れてたい	ニッポン	日本	名詞-固有名詞-地名-国	
	OC04_00001_c	530	帰国ということなのですが、なんだかうそ臭くて。。。	日本	国内ということなら医学部医学科卒業は早くても二十四	ニッポン	日本	名詞-固有名詞-地名-国	
	OC05_00327_c	490	た週は、あるので、しょうか？彼らの週は彼らが作る。	日本	人がどうこう言える問題じゃない。	ニッポン	日本	名詞-固有名詞-地名-国	
	OC06_02180_c	800	七百万円」となっています。その方に今かったのが、	日本	人では山田久志の三千九百万円だから、倍以上ですね。	ニッポン	日本	名詞-固有名詞-地名-国	
	OC09_02741_c	280	n？以前どこかで骨量測定？のようなものをやったら	日本	人平均女性の1．2倍ぐらいありました。百六十五cmあ	ニッポン	日本	名詞-固有名詞-地名-国	
	OC12_01186_c	750	が、英文はなかなか見つからなくて。お願いします。	日本	語と同じで、必ずしもこれでなければダメということはない	ニッポン	日本	名詞-固有名詞-地名-国	
	OC14_03063_c	1020	少も、やはり「浄やかん」時に見られるのかなたいたい	日本	語が使えない人がここを見る可能性も低いも。	ニッポン	日本	名詞-固有名詞-地名-国	

レコード 1 / 10　検索

図 41　全文検索による検索結果の例

高度な検索は 3 語までの繋がりについて検索することができる。検索項目はコーパスデータベースのほぼ全ての項目を網羅していて、さらに辞書データベース等の項目も指定することができる。たとえば辞書データベースの短単位語彙素テーブルの値に対して検索条件を指定するような複雑な式を記述することも可能である。また、高度な検索の条件式は保存することができるので、管理者が複雑な検索条件を作成して保存すれば、作業者が同じ条件で検索することが可能である。

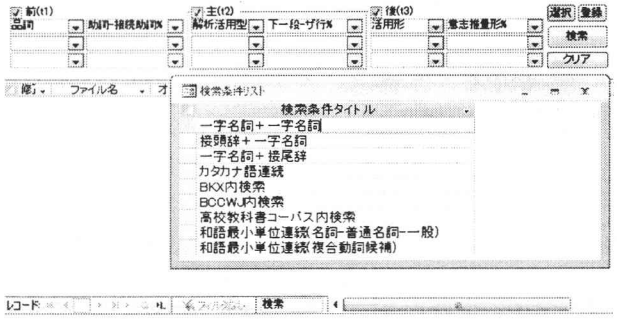


図 42　高度な検索の例

高度な検索結果の表示例を以下に示す。高度な検索は、特定の語の繋がりのパターンを検索したい場合などに有効である。

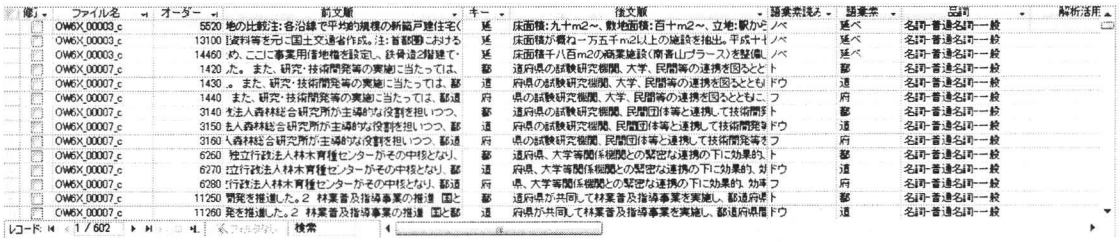


図 43　高度な検索による検索結果の例

検索は、各検索方法専用のストアドプロシージャで処理される。各ストアドプロシージャは、独自のロジックで短単位テーブル内の検索を行うが、検索結果が作業者専用の作業テーブル内に保存されるという点で共通している。各検索ストアドプロシージャが独立していることによって、検索の機能拡張や修正などを容易に行うことができる。

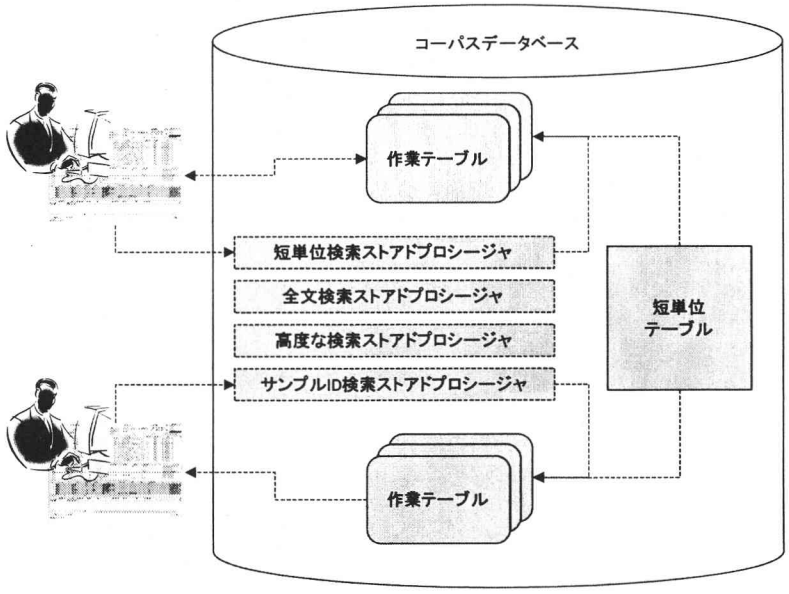


図 44　検索用ストアドプロシージャと作業テーブル他の関連図

6.4.2.検索対象コーパスの指定

コーパスデータベース（の短単位テーブル）には1億語が格納されることを想定しているが、日常的な作業時にデータベース全体に対して検索や更新を行うことは殆どない。にもかかわらず、検索時にデータベース全体を検索対象にしてしまうと、検索にかかる負荷が増大してしまい、作業効率が低下してしまう。そこで、大納言では前述の 4 種類以外の検索条件として、検索対象コーパスを指定することができるようにしている。

検索対象コーパスの指定は、前述の 4 種類の検索方法と組み合わせて使用する。また、検索対象コーパスは複数指定することもできる。たとえば、白書コアデータに対する出現書字形の検索や、書籍コアデータと新聞コアデータに対する全文検索をすることができる。

検索対象コーパスを指定することによるメリットとしては、検索対象を絞ることによる検索時の負荷の低減がある。また、作業（検索）対象を制限できるので、作業者の意図しないコーパスのデータ変更を防ぐメリットもある。

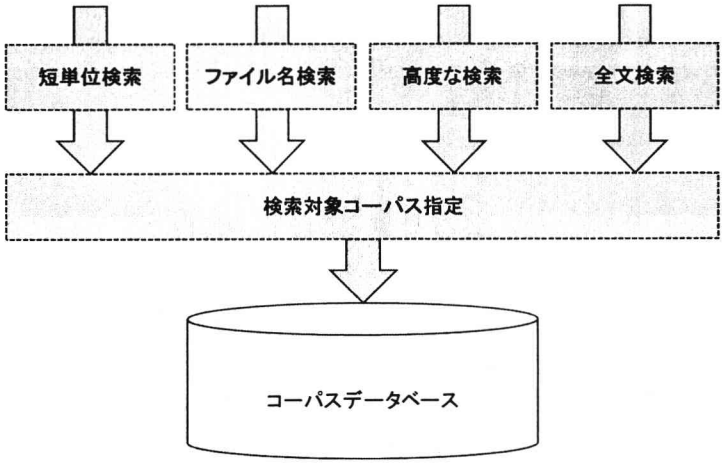


図 45　検索方法指定の概念図

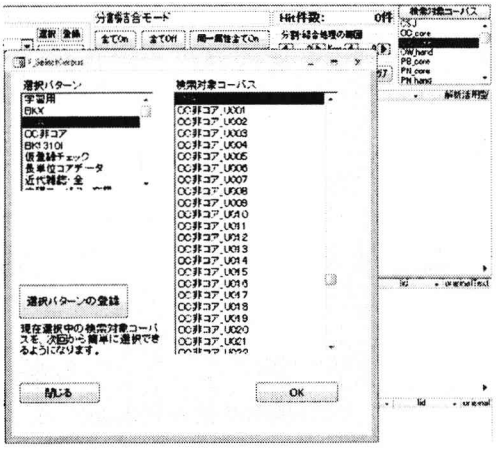


図 46　検索対象コーパスの指定の例

6.4.3.前後文脈生成処理

KWIC 画面では語についての前後文脈が表示されるが、コーパスデータベース内には語についての前後文脈を格納していない。なぜなら、コーパスデータベースは総語数 1 億語を想定している為に、その語の全てについて文脈を格納するというのは、データベースの容量上も、管理上も適切ではないからである。また、全ての語について前後文脈を格納すると、文字修正処理や対話式数字変換処理のような出現書字形が変更される処理の際に、実際の修正レコード以外も更新する必要があり、処理の負荷が増大してしまうからである。

以上のようなことを考慮して、大納言では検索の都度、短単位テーブルの出現書字形から文脈を生成する処理を行うことで、前後文脈を取得している。

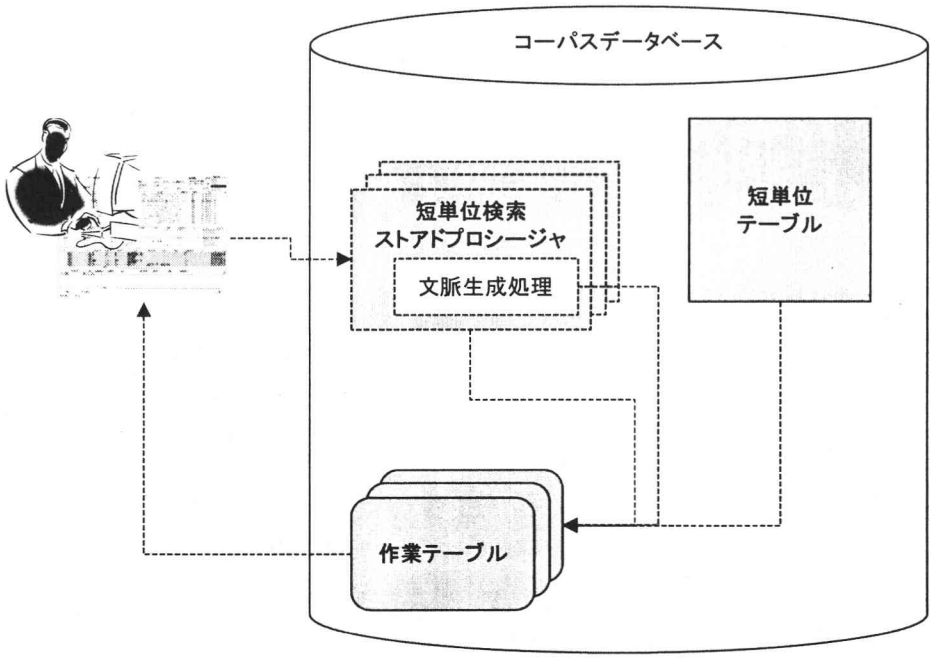


図 47　文脈生成処理概念図

ただし、検索のたびに文脈を生成するということは、文脈を生成する処理の分だけ検索結果の取得に時間がかかるというデメリットがある。このデメリットを可能な限り少なくするために、短単位テーブルはサンプル ID と連番にクラスタ化インデックスを設定している。これによって、語の出現順とデータの物理的な順序関係が一致し、文脈生成時の短単位テーブルの並べ替えを不要にすることができる。

問題は連番の振り方である。もし連番が 1、2、3…と隙間なく振られていた場合、連番が 2 の語を分割処理しようとすると、連番が詰まっているために、追加するレコードに連番が振れなくなってしまう。こうした点を考慮して、短単位テーブルの連番を 10、20、30…の

ように 10 間隔で振り、分割結合時に追加するレコードには端数（10 で割り切れない数)を振ることによって、新規レコードを既存レコード間に挿入できるようにしている。

分割結合時の具体的な連番の振り方の例を示す。連番=10、出現書字形「これは」を「これ」と「は」に分割する場合、修正する先頭の語の連番を n とすると、修正後の語に振られる連番を n+1、n+2…のように端数にする。こうすることで、語の物理的な相対位置を変えずに新規レコードを挿入することができる。

なお、これによって生じた連番の端数は、定期的に実行されるジョブ処理（連番振り直し処理)によって解消される。また、レコードの挿入によってインデックスページの断片化が起こらないよう、インデックスの構築時にインデックスページ内にあらかじめ空き領域を設けている。

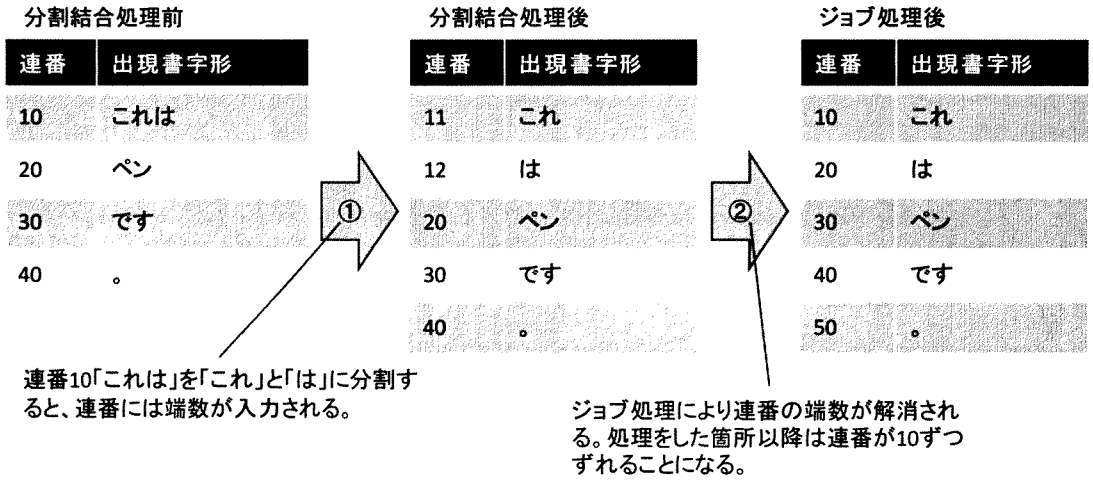


図 48　分割結合処理・ジョブ処理時の連番の振り方

短単位テーブルの連番の端数は、データの整合性維持にも利用されている。

たとえば、複数の作業者（A・B)がいる場合に、作業者 A が作業テーブルにデータを読み込んだ後に、同じ箇所を作業者 B が更新したとする。通常、複数の作業者による同一レコードの修正はデータの不整合を引き起こす原因になることが多いが、大納言では作業者 A が更新する際には短単位テーブルに該当するレコードが存在しない（作業者 B による更新によって連番が変更されている)場合には、作業者 A の処理はキャンセルされるようになっている（図 49 参照）。

なお、文脈生成処理は短単位検索処理以外（サンプル ID 検索・全文検索・高度な検索時）でも同様である。各検索プログラムは内部に文脈生成処理を含んでいて、短単位テーブルから該当データを検索し、文脈を生成した後に作業テーブルに格納している。

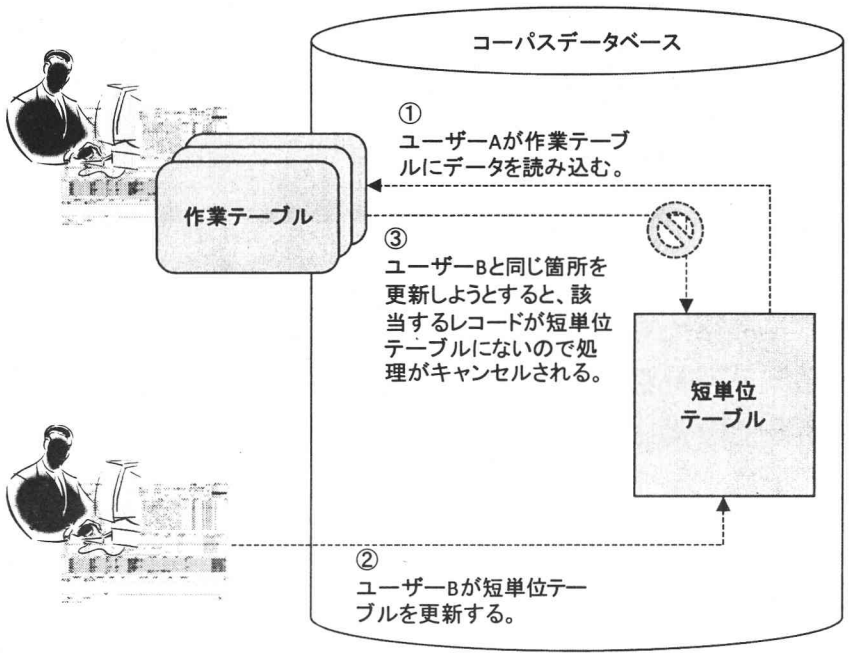


図 49　連番の端数によるデータ整合性維持

6.4.4.全文検索機能

全文検索は、単純に短単位テーブルのみを使用して処理を行おうとすると短単位境界を越えて検索することになるので、データベースに負荷がかかってしまう。また、全文検索用のシステムで通常用いられる転置インデックスは 1 億語規模のコーパスデータベースではインデックスのサイズが巨大になってしまうため適切ではない。そこで、大納言では SQL Server の全文検索機能を利用した独自の全文検索処理を行っている。

大納言の全文検索の仕組みでは、全文検索用の文テーブルを使用している。文テーブルにはサンプル名と文と、そのサンプル内での文の開始位置が格納されている。一方短単位テーブルには文テーブルと対応する形でサンプル内での語の開始位置が格納されている（表 22・表 23）。

全文検索の処理の流れは以下の通りである（図 50 参照）。作業者が大納言を使用して全文検索を実行すると、検索文字列を受け取った全文検索プログラムは一次処理として文テーブルに対して文字列の検索を行い、該当する文字列を含むレコードのサンプル ID と文開始位置、その文中における検索文字列の出現頻度を求め、一次検索結果テーブルに格納する。次に二次処理として、一次処理結果で出現頻度が 1 のレコードについて、詳細な文開始位置を求め、二次検索結果テーブルに格納する。更に三次処理で、一次処理結果で出現頻度が 2 以上のレコードについて、文中に存在する検索文字列の全ての詳細な文開始位置

を求め、三次検索結果テーブルに格納する。こうして調べられた文開始位置について短単位テーブルを検索し、その結果を作業テーブルに格納する。

表 22　短単位テーブルと文テーブルのデータ例（短単位テーブル）

サンプル ID	文境界	出現書字形	文開始位置	文終了位置
OW6X_00000	B	1	10	20
OW6X_00000	I		20	30
OW6X_00000	I	日本	30	50
OW6X_00000	I	文化	50	70
…	…	…	…	…
OW6X_00000	B	（	220	230
OW6X_00000	I	1	230	240
OW6X_00000	I	）	240	250
…	…	…	…	…
OW6X_00000	B	1	350	360
OW6X_00000	I		360	370
OW6X_00000	I	文化	370	390
OW6X_00000	I	庁	390	400
…	…	…	…	…

表 23　短単位テーブルと文テーブルのデータ例（文テーブル）

サンプル ID	文開始位置	文
OW6X_00000	10	1　日本文化の発信による国際文化交流の推進
OW6X_00000	220	（1）文化庁文化交流使事業
OW6X_00000	350	1　文化庁文化交流使事業
…	…	…

なお、文字修正処理や数値変換処理によって本文が変更された場合には、文テーブルの該当箇所も変更する必要があるが、この処理はジョブによって行われる。ジョブ処理では文テーブルと短単位テーブルの間の不整合を検出し、整合性を維持するようそれぞれのテーブルを毎日自動的に更新している。

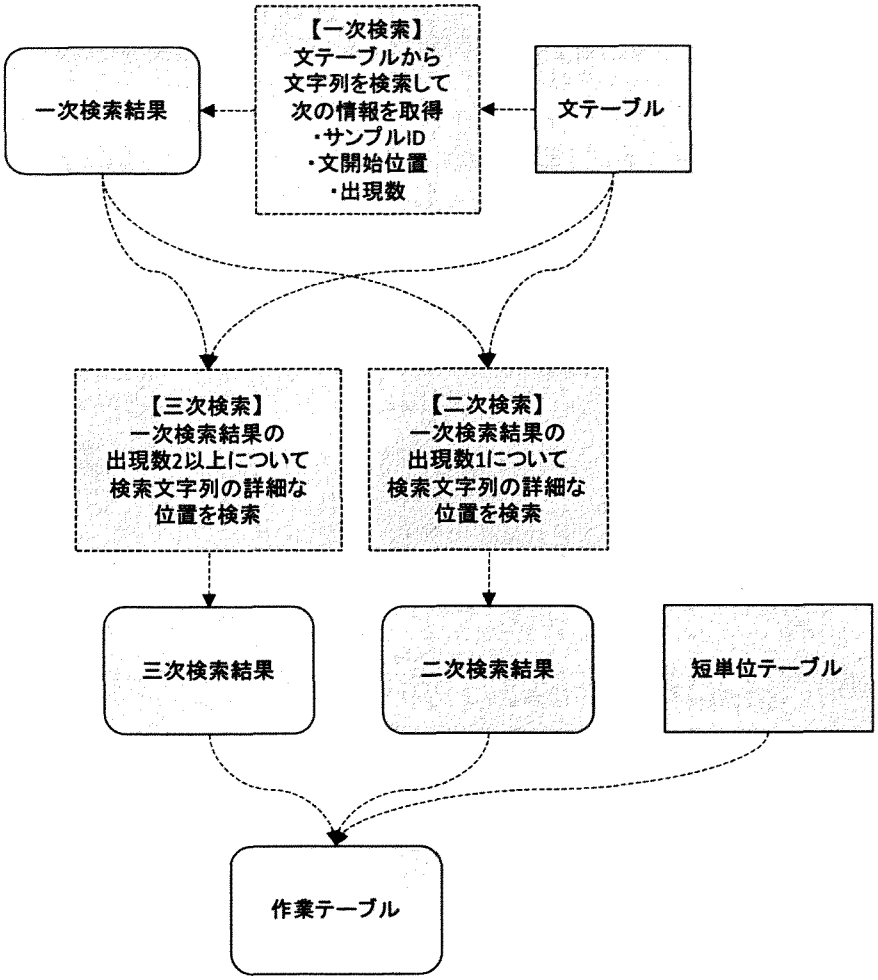


図 50　全文検索処理の概念図

6.5. 分割結合処理

6.5.1. 分割結合処理の概要

分割結合処理は語の区切り位置を修正して、さらに語に属性を付与するための処理である。

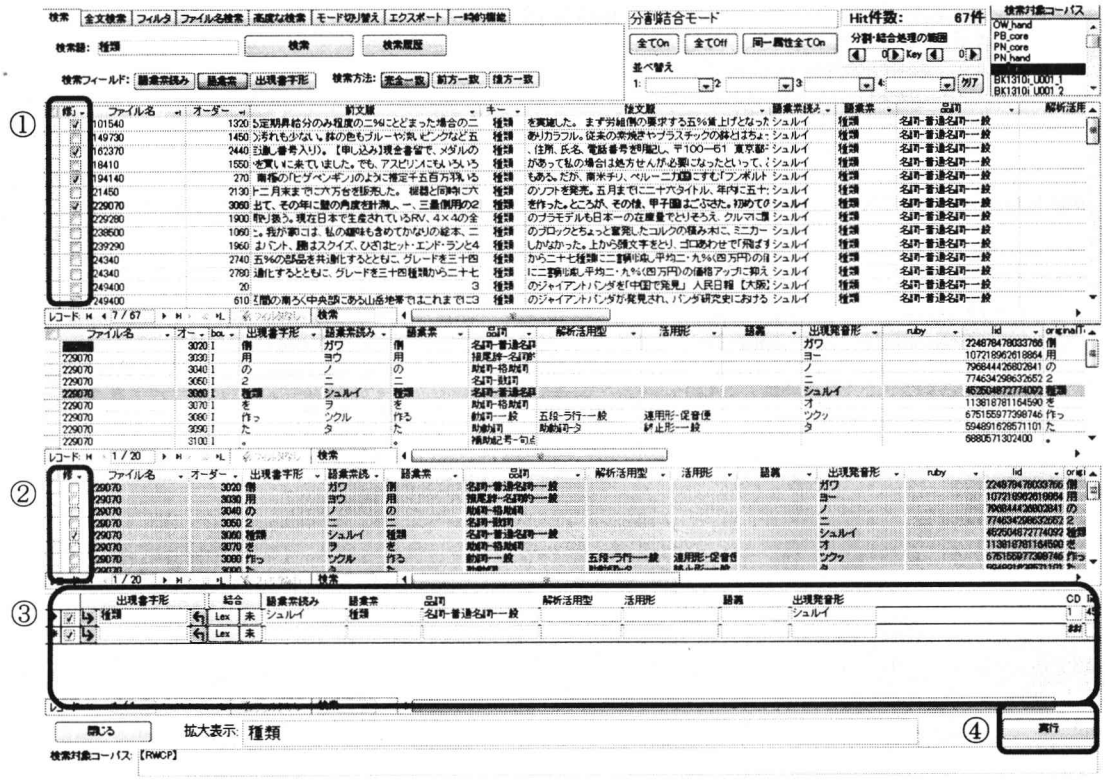


図 51　分割結合処理時の操作方法

大納言における短単位修正時の操作方法は、修正する語を KWIC サブフォームから選択し（①）、修正する範囲（前後の範囲）を指定し（②）、語の区切りの修正と正しい属性の付与をし（③）、実行ボタンをクリックすることによりデータベースに反映する（④）。

なお、③における語の属性の付与は語彙表テーブルの中から適切なものを選択することで行う。これによって、短単位テーブルと語彙表とが関連付けされ、辞書データベースとも関連付けられることになる。辞書データベースに関連付けられた短単位については、ジョブ処理によって整合性が維持される。また辞書データベースで行った変更は、ジョブ処理によって語彙表テーブルを介して短単位テーブルにも反映される。

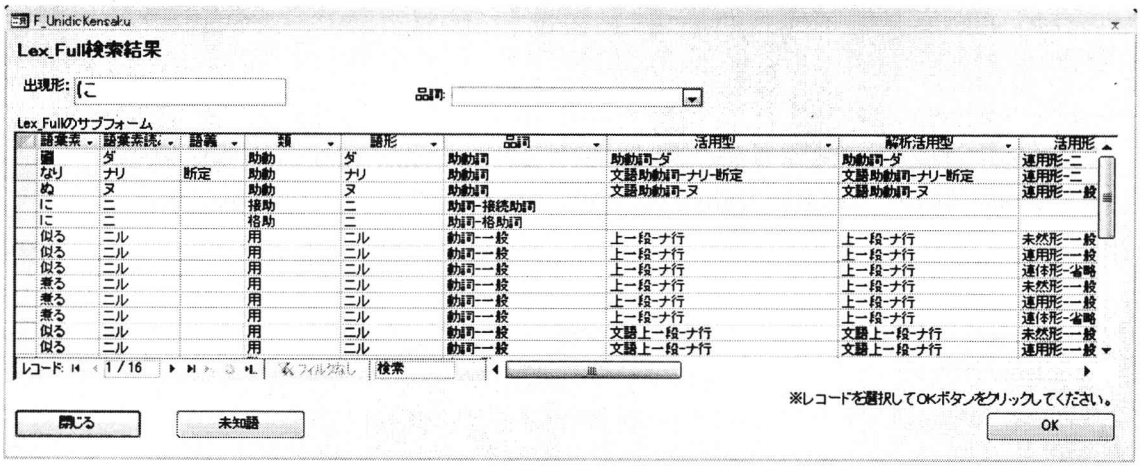


図 52　語彙表テーブル参照用画面

6.5.2.分割結合時のデータチェック機能一覧

大納言では、分割結合処理時に各種のデータチェックを行っており、データに不整合が起こらないようにしている。データチェックの種類と詳細は下記の通りである。

表 24　分割結合時のデータチェック機能

名称	チェック内容	タイミング	適用されるモード
同一属性チェック	大納言では同一属性を持つ語を一括で処理をすることができる。逆にいうと、同一属性でない語は一括処理できない。同一属性チェックは、処理しようとしている複数の語が同じ属性値であるかを調査する処理である。同一属性チェックを行う項目は以下の通りで、これらの項目が同じ値を格納していれば、一括処理を行うことができる。 ・出現書字形 ・出現発音形 ・品詞 ・活用型 ・活用形 ・語彙素読み ・語彙素 ・語彙素細分類	ツール操作時	短単位 長単位 数字変換処理
文境界チェック	文境界を越えて処理することはできない。	ツール操作時	短単位 長単位 数字変換処理

連番端数チェック	連番が 10 の倍数でないものは処理することはできない。	ツール操作時	短単位 長単位 数字変換処理 文字修正処理
数字タグ境界チェック	数字タグ境界を越えて処理することはできない。	ツール操作時	短単位 長単位 文字修正
数字タグ範囲チェック	数字タグ範囲内は処理できない。	ツール操作時	文字修正
文脈整合性チェック 1	作業テーブルにおいて修正前と修正後の文脈の相違をチェック。	ツール操作時	短単位 長単位
文脈整合性チェック 2	作業テーブルと短単位テーブルの文脈の相違をチェック。	分割結合処理時	短単位 長単位
文脈整合性チェック 3	実際に処理を行った結果について、処理前後の文脈の相違をチェック。	分割結合処理時	短単位 長単位

6.5.3.同一属性レコードの一括処理

大納言では、同じ属性値を持つ複数の語については、一括処理をすることができる。また一括処理に関する作業を補助する機能も実装している。以下に一括処理の例を示す。なお、同一属性チェックを行う項目は出現書字形・出現発音形・品詞・活用型・活用形・語彙素読み・語彙素・語彙素細分類である。

単純な同一属性一括処理例

誤った語の属性

サンプル ID	順番	出現書字形	出現発音形	(その他の属性)
A001	10	国語	A	B
…	…	…	…	…
A002	150	国語	A	B
…	…	…	…	…
A003	980	国語	A	B

正しい語の属性

出現書字形	出現発音形	(その他の属性)
国語	C	D

↓一括処理

サンプル ID	順番	出現書字形	出現発音形	(その他の属性)
A001	10	国語	C	D
…	…	…	…	…
A002	150	国語	C	D
…	…	…	…	…
A003	980	国語	C	D

複雑な同一属性一括処理パターン例

誤った語の属性

サンプル ID	順番	出現書字形	出現発音形	(その他の属性)
A001	10	書	A	B
A001	20	字形	C	D
…	…	…	…	…
A002	90	書	A	B
A002	100	字形	C	D
…	…	…	…	…
A003	5300	書	A	B
A003	5310	字形	C	D
…	…	…	…	…

正しい語の属性

出現書字形	出現発音形	(その他の属性)
書字	E	F
形	G	H

↓一括処理

サンプル ID	順番	出現書字形	出現発音形	(その他の属性)
A001	11	書字	E	F
A001	12	形	G	H
…	…	…	…	…
A002	91	書字	E	F
A002	92	形	G	H
…	…	…	…	…
A003	5301	書字	E	F
A003	5302	形	G	H
…	…	…	…	…

同一属性の一括選択は、フォーム上のボタンをクリックすることで行う。このボタンにより、KWIC で作業者が選択中のものと同じ属性（前後の処理範囲の語の属性まで同じもの）を持つものを自動で選択することができる。

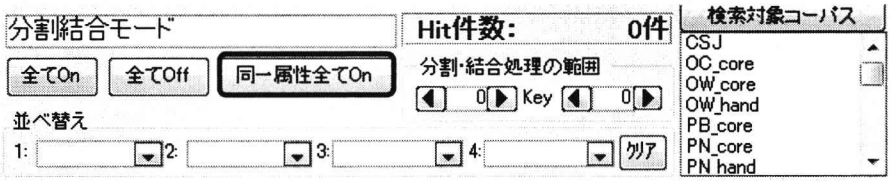


図 53　同一属性レコードの一括選択ボタン

6.5.4.文字位置取得処理

短単位テーブルを更新する場合には、文字テーブルとの間でサンプル ID、文字開始位置、文字終了位置の対応関係を保つ必要がある。複数の短単位を一括処理する場合や、短単位が文字修正（文字開始位置・終了位置に端数を格納している）されている場合も同様である。このように処理時に短単位テーブルと文字テーブルの対応をとるための処理が文字位置取得処理である。

文字位置取得処理は短単位テーブル更新処理時に呼び出される。文字位置取得処理は文字テーブルを参照して作業用テーブルに文字開始位置・終了位置を入力する。短単位テーブルを更新するストアドプロシージャはこの作業用テーブルを利用して短単位テーブルを更新する（図 54）。

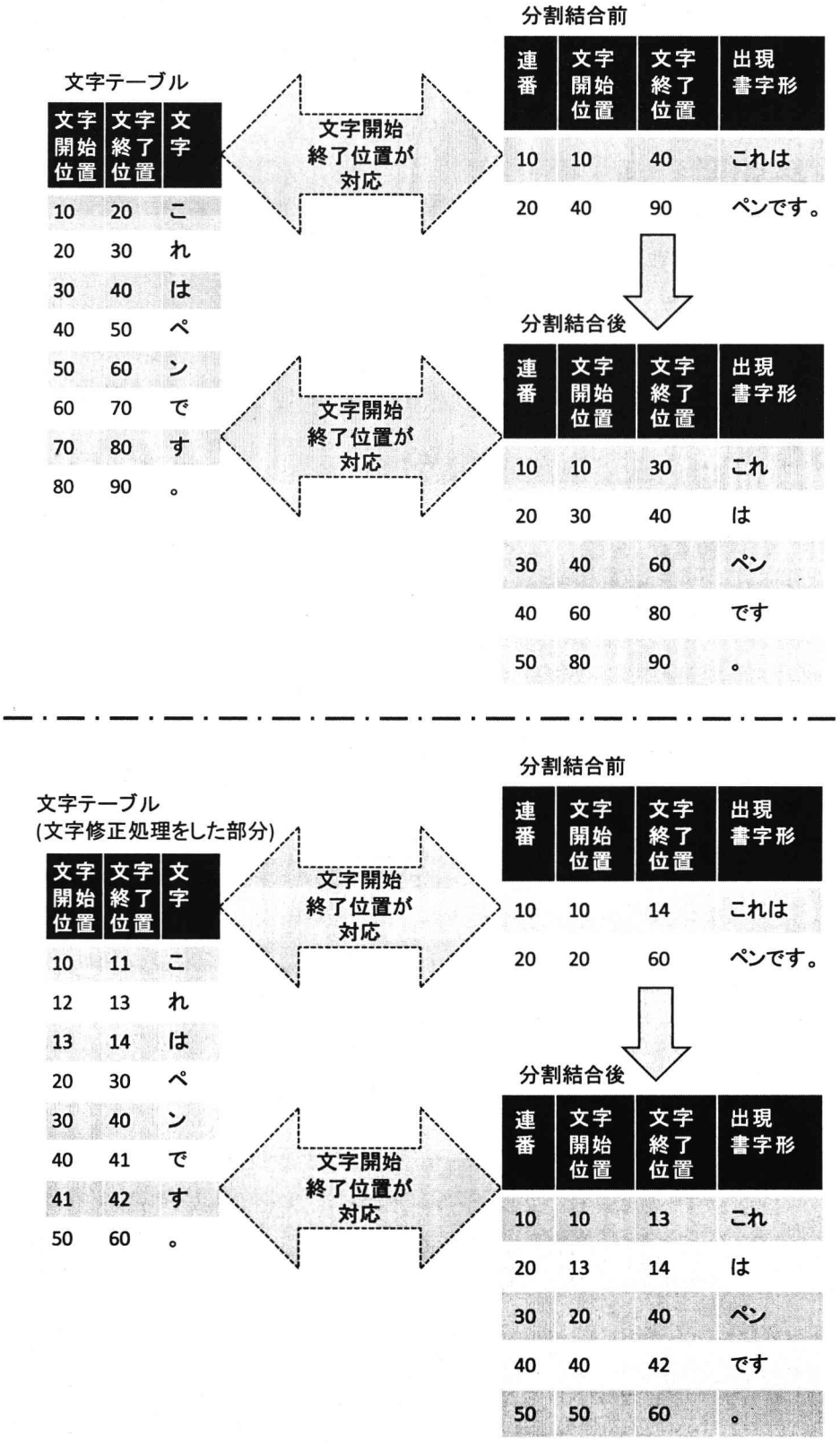


図 54　文字位置取得処理

6.5.5.文脈チェック処理

大納言では複数の作業者に同時に利用されることを想定しているが、同時実行性を高めるためにレコードロックは必要最小限にとどめている。ただしこの方法は複数の作業者により同一箇所が更新された場合に、文脈の整合性が維持されないリスクがある。そのため、大納言では短単位テーブル更新処理の際に何重もの文脈チェック処理を行うことで、文脈が崩れないようにしている。

分割結合処理内で行う文脈チェック処理については、作業テーブル内文脈整合性チェック、作業テーブル短単位テーブル文脈整合性チェックと、処理前後文脈整合性チェックの3種類ある。

作業テーブル内文脈整合性チェック

最初に行われる作業テーブル内文脈整合性チェックは、作業テーブル内に読みこんだ短単位について、修正前と修正後（大納言入力後、短単位テーブルに反映する前）の文脈の整合性をチェックする処理である。これは、操作上のミスやツールの問題などによって起こる文脈の変更を防ぐために行っている。これは大納言での操作中に行われる処理なので、チェックを通過できない場合は短単位テーブル更新処理が実行できなくなっている。

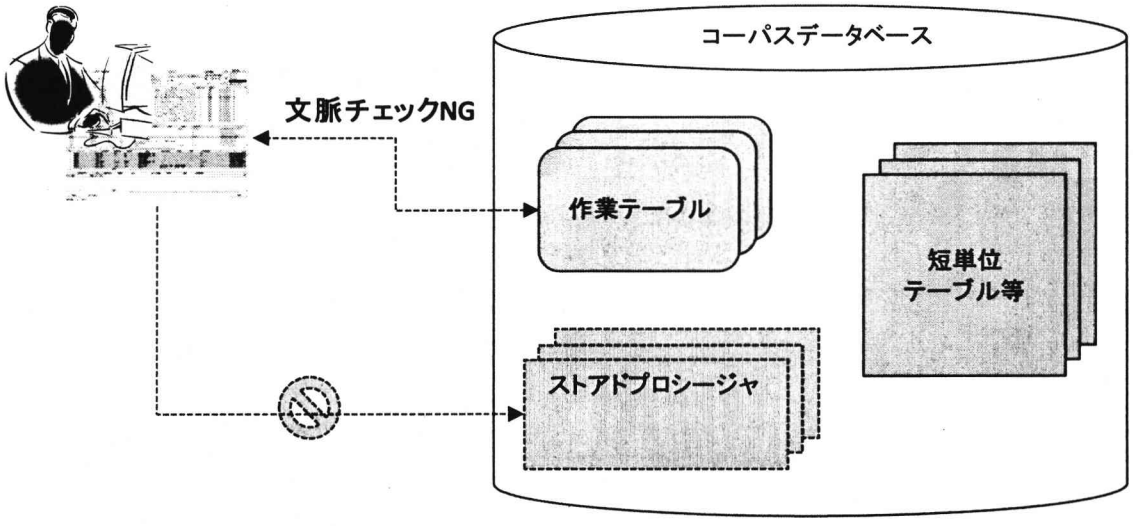


図 55　作業テーブル内文脈整合性チェック

作業テーブル短単位テーブル文脈整合性チェック

短単位テーブルに対する更新処理中に行われる作業テーブル短単位テーブル文脈整合性チェックでは、作業テーブルの内容と短単位テーブルの内容の整合性がチェックされる。これは主に複数の作業者が短単位テーブルを更新することによって文脈が崩れることを防ぐために行われるものである。

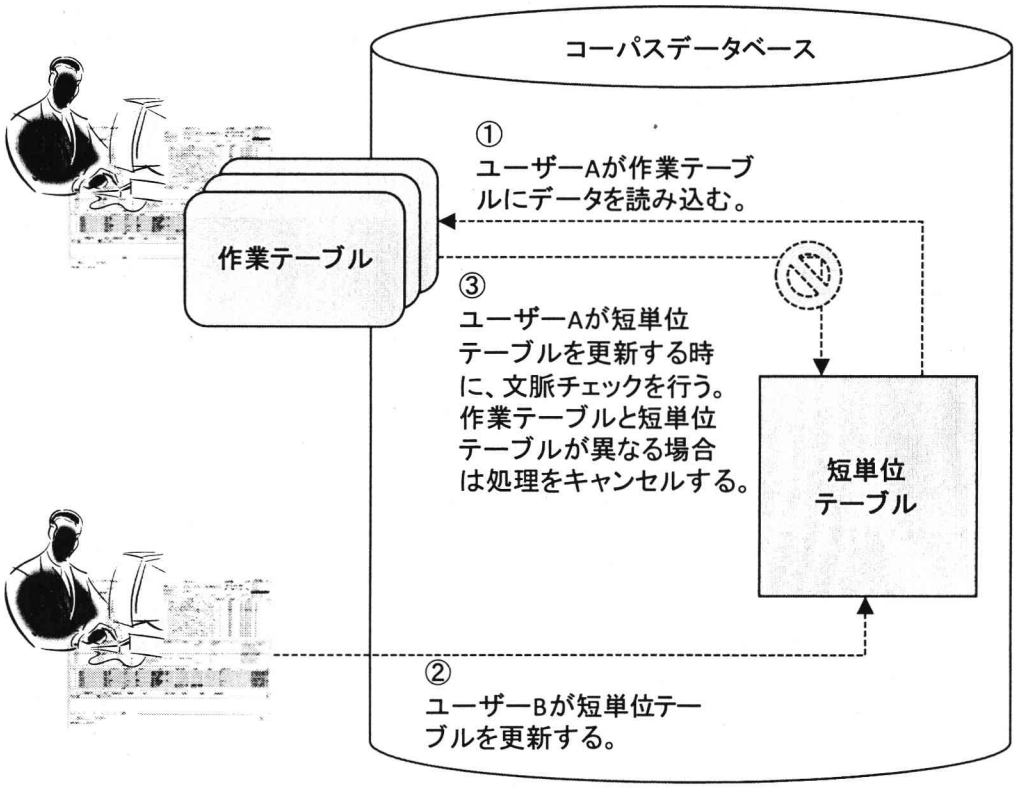


図 56　作業テーブルと短単位テーブル間の文脈整合性チェック

処理前後文脈整合性チェック

処理前後文脈整合性チェックはトランザクションで括られた処理の先頭と最後で、文脈の比較をする処理である。トランザクションで括られた範囲には短単位テーブル更新処理以外にもいくつかの処理が含まれるため、わずかとはいえ、トランザクション処理中に他の作業者により短単位テーブルが更新される可能性があり、そのまま処理してしまうと文脈が崩れてしまう危険がある。それを回避するための処理が処理前後文脈整合性チェックである。

トランザクション処理中の文脈の整合性を維持するために考えられる他の方法としては、トランザクションの分離レベルを設定するという方法あるが、この方法は同時実行性が低下するため、複数作業者を想定している大納言においては作業性の点から不適切である。そのため、大納言では文脈チェック処理を行うことで、同時実行性と文脈整合性の維持を両立させている。

なお、処理開始レコードの前 1 レコードから処理開始レコードの後 1 レコードまでを文脈チェック対象レコードとしている。

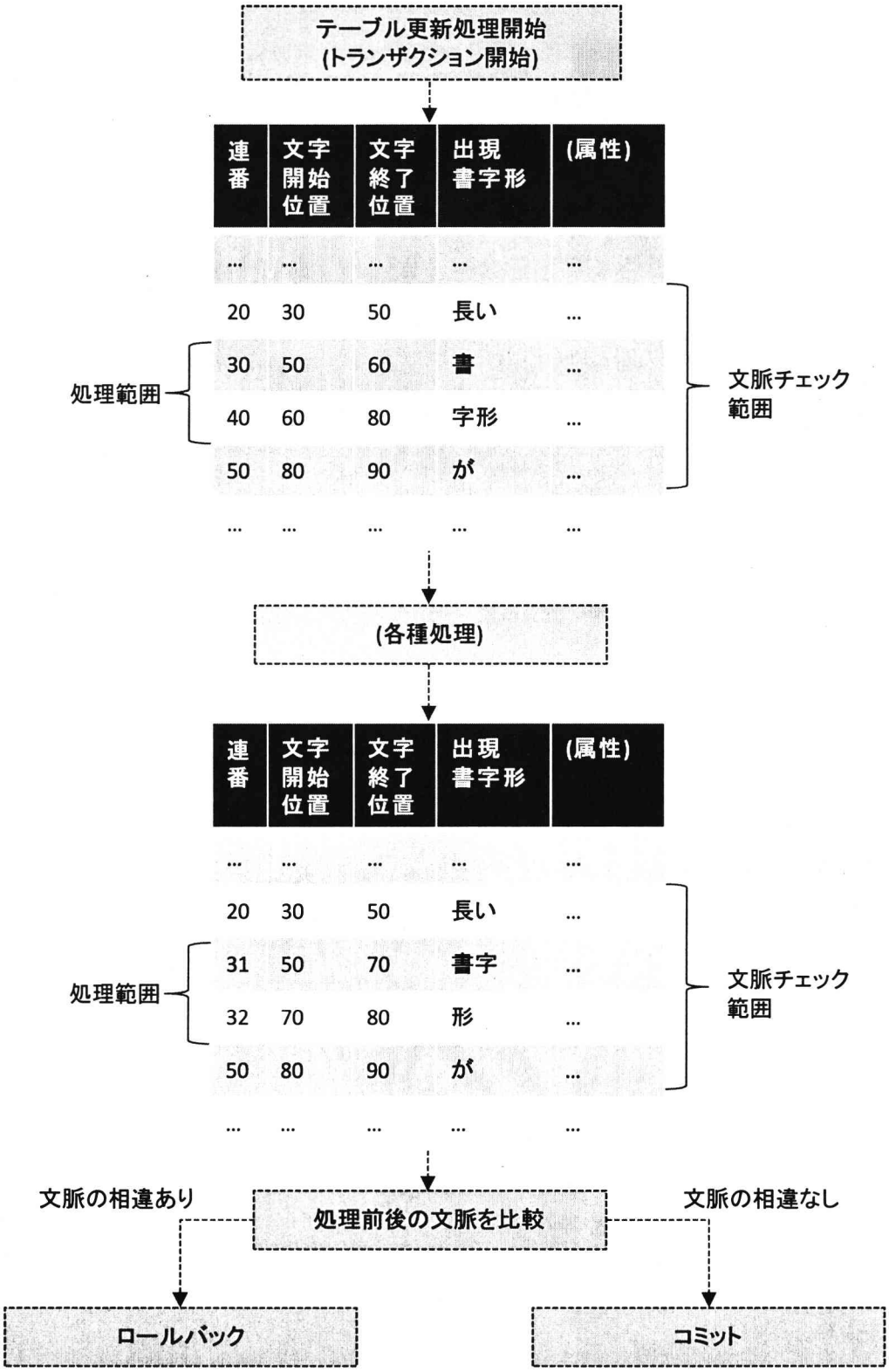


図 57　処理前後文脈整合性チェック

ただし、対話式数字変換処理・文字修正処理時には、処理前後文脈チェックは行わない。これらの処理は文脈が変更される処理だからである。対話式数字変換処理と文字修正処理時は文脈確認用画面を表示して、作業者の目視により文脈の整合性を確認するようにしている。

図 58　目視による文脈の確認画面

これら文脈チェック処理や文字位置取得処理の流れをまとめたものが以下の図である。

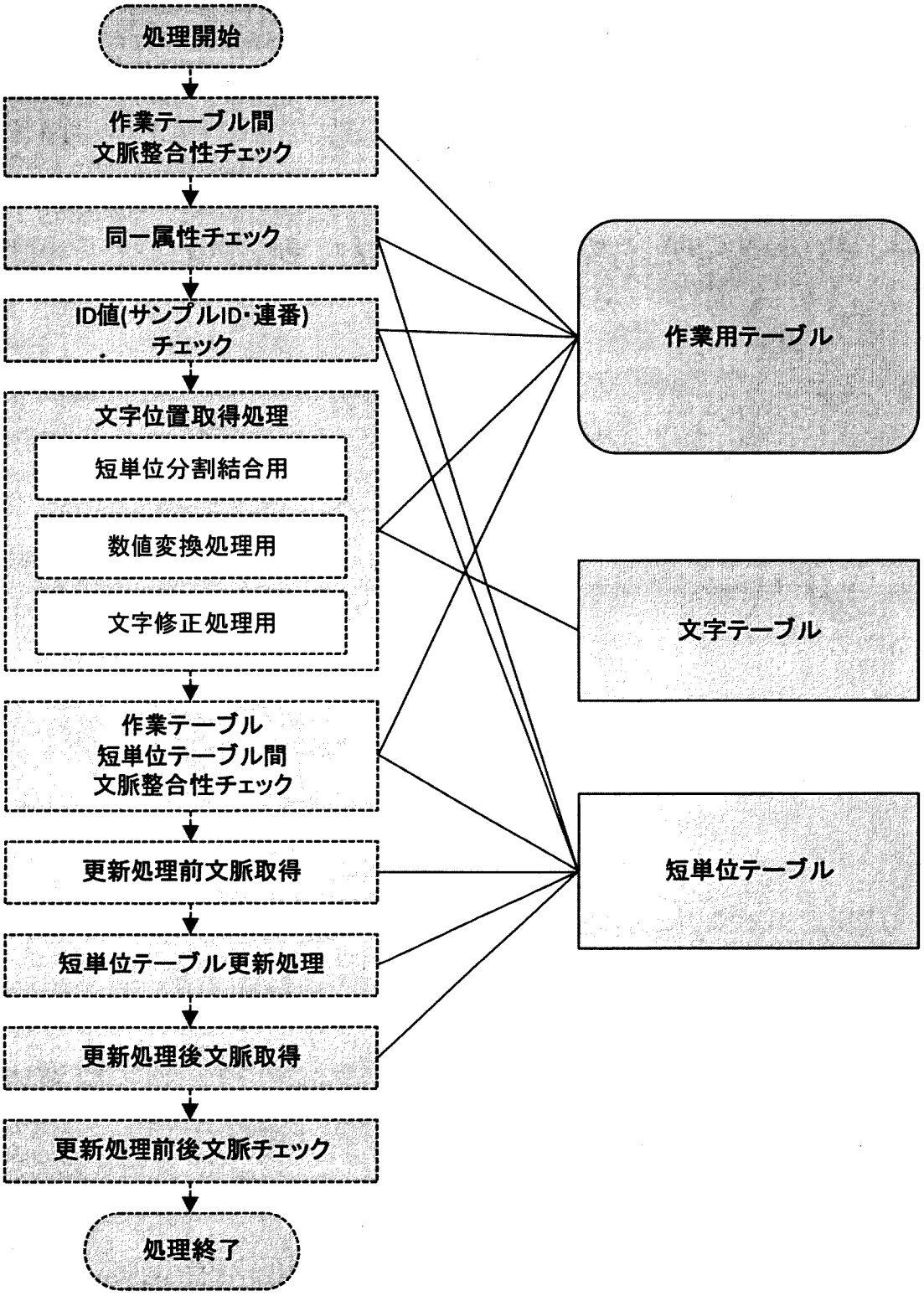


図 59　短単位テーブル更新処理の流れ

これらの処理が全て通って初めて短単位テーブルの更新が確定される。いずれかのプロセスで問題が検出された場合は、処理はキャンセルまたはロールバックされる。また、対話式数字変換処理時・文字修正処理時にはそれぞれ専用の文字位置取得処理が行われる。

6.5.6.短単位テーブル更新時の長単位テーブル更新処理

短単位テーブルの更新が長単位の境界をまたぐ場合は、長単位テーブルの該当箇所の長単位の区切りと属性を見直す必要があるため、短単位テーブル更新時に長単位テーブルに及ぼす影響をチェックして、必要であれば長単位テーブルの該当箇所にフラグをたてる処理を行っている。作業者はフラグを検索することで短単位境界と長単位境界の相違を容易にチェックすることができる。

6.5.7.特殊な属性値

分割結合作業における属性付与時に、語彙表には存在しない特殊な属性値を付与することができる。特殊な属性値は以下の通りである。

表 25　主な特殊属性値

ID	属性値	説明
1	新規未知語	一致するものが語彙表内に存在しない語
2	英単語	辞書登録を行わないアルファベット表記の語
3	電子化誤り	（作業用）BCCWJ の電子化の際の誤り
4	コンピュータ用語	辞書登録を行わないコンピュータ用語（関数名等）
6	correct 処理	（作業用）原文修正処理を行った箇所
7	URL	URL、メールアドレス等（解析を行わない）
8	電子化ママ	（作業用）BCCWJ の電子化の際の不審箇所
9	漢文	サンプル中の漢文（解析を行わない）
10	方言	サンプル中の方言会話（解析を行わない）
11	振り仮名	（作業用）本文中に陥入する括弧入りの振り仮名
12	チェック済み	（作業用）
13	NumTrans 処理	（作業用）数字処理を行った箇所
14	カタカナ文	（作業用）サンプル中のカタカナ漢字交じり文
15	言いよどみ	辞書登録を行わないサンプル中のいいよどみ
16	web 誤脱	Web データ特有の誤脱

特殊な属性値が付与された語については、高度な検索を利用して検索することができる。

図 60　高度な検索による特殊な属性値の検索例

6.6. 対話式数字変換処理

6.6.1. 対話式数字変換処理の概要

UniDic での解析において、アラビア数字で書かれた本文を漢数字に変換する等の数字変換（NumTrans）処理が行われる。形態論情報データベースに取り込まれたデータを修正する際、この数字変換処理の誤りを手動で直したり、数字変換処理が為されなかった部分に手動で変換処理を行ったりする必要が生じる。このための機能が大納言の対話式数字変換処理機能（モード）である。対話式数字変換処理モードでは、アラビア数字で書かれた本文を漢数字や分数などに変換するための操作をサポートする。

なお、対話式数字変換処理では次のような短単位分割結合時にはない処理が行われる。

・出現書字形が変更される。
・文字開始位置と文字終了位置が通常とは異なる形で振られる。
・短単位テーブルの他に、数字テーブルと文テーブルが更新される。

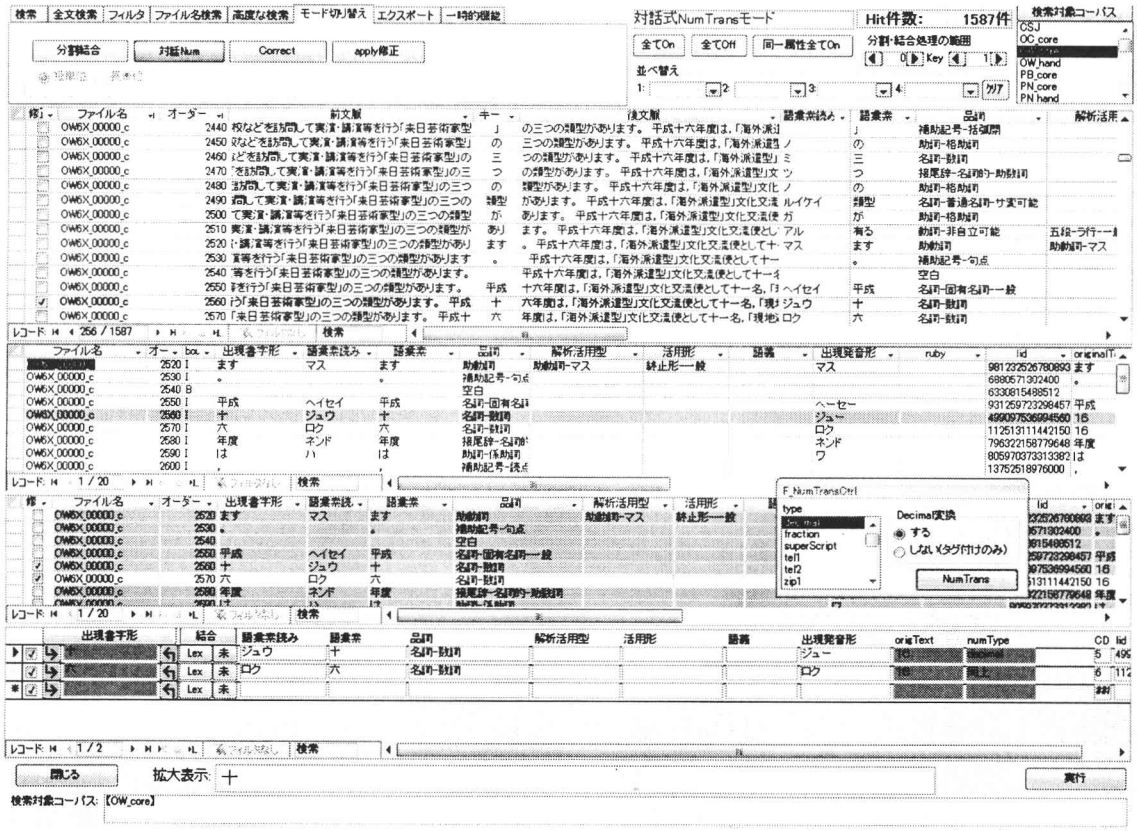


図 61　対話式数字変換処理の作業画面

6.6.2.数字変換処理の種類

対話式数字変換処理の種類には以下のものがある。

表 26　数字変換処理の型

変換型	説明	変換例
Decimal 変換	一般の数字の変換	１９９７ → 千｜九百｜九十｜七
Fraction 変換	分数の変換 （BCCWJ の fraction タグ）	１／２ <fraction> １ ／ ２ </fraction> → ２｜分｜１
SuperScript 変換	上付き数字の変換 （BCCWJ の superScript タグ）	2^3 ２<superScript> ３ </superScript> → ２｜３｜乗

※　NumTrans による数字変換を経た場合には fraction タグの仕様が異なる。詳細については NumTrans のマニュアルを参照のこと。

6.6.3.テーブル間の整合性について

対話式数字変換処理をした際は、短単位テーブル以外のテーブルも更新し、関連する各テーブル間に矛盾が起こらないようにしている。

まず、対話式数字変換処理によって短単位テーブルが更新され、次に数字タグ情報が数字テーブルに保存される。また、対話式数字変換処理は短単位の出現書字形が変更される処理なので、長単位テーブルも更新される。

さらに、出現書字形が変更されるということは、文開始位置・文終了位置も変更されることになるので、短単位テーブルの文開始位置・終了位置と文テーブルも更新される。ただしこの処理はリアルタイムではなくジョブ処理により行われる。

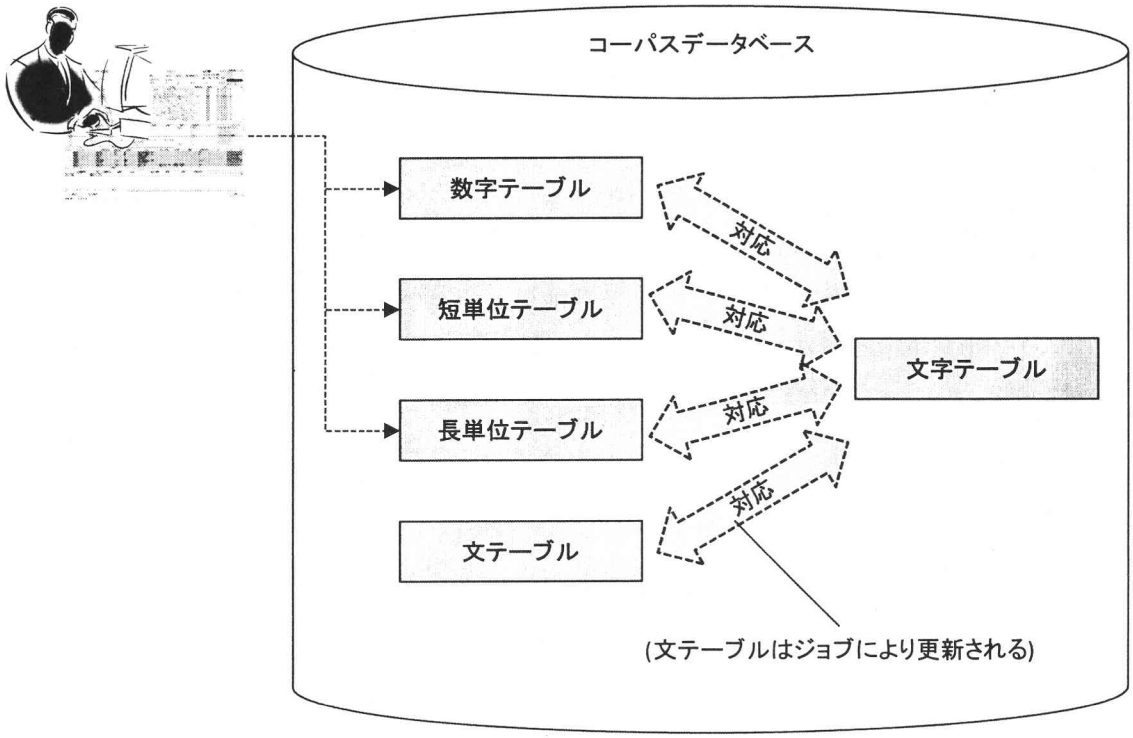


図 62　対話式数字変換時の各テーブルの対応関係

6.7.文字修正処理

6.7.1.文字修正処理の概要

文字修正処理は、文字テーブルにある文字を別の文字に変更したり、文字の追加・削除をするための処理である。大納言では文字修正モードに切り替えることで文字修正機能が利用できる。

図 63　文字修正処理の作業画面

6.7.2.文字修正処理の種類

文字の修正型の種類には表 27 に示すものがある。文字修正した際には、文字修正テーブルに、修正箇所などとともに記録される。

文字修正の記録は BCCWJ の correction タグに相当するものであり、XML 出力時には correct タグとして出力される。

表 27　文字修正処理の種類

型	説明
誤字	文字の誤り
脱字	文字の脱落
衍字	余分な文字の挿入
誤変換	誤変換による単語単位での誤字

6.7.3.テーブル間の整合性について

文字修正処理における文字の追加・変更・削除は、対応する短単位テーブル、長単位テーブル、文テーブル等にも影響を与えるため、これらのテーブルも更新する必要がある。

また、文字修正によって文字開始・終了位置が変更されることもあるため、この場合にもテーブル間の対応がとれるように文字開始・終了位置を更新する必要がある。文字修正処理はこれらの対応が維持されるよう行われる。また処理の単純化と作業時のミスを避けるために、同一属性一括処理には対応していない。

なお、図 64 にて数字テーブルが処理対象に含まれていないのは、対応するレコードを数字テーブルに持つ短単位についての文字修正は、大納言で許可しないようにしているからである。このような部分について文字修正処理をする場合は、対応するレコードを数字テーブルから削除して、該当部分の数字テーブルと短単位テーブルの連動を解除する必要がある。連動の解除は大納言の対話式数字変換処理を利用して行う。

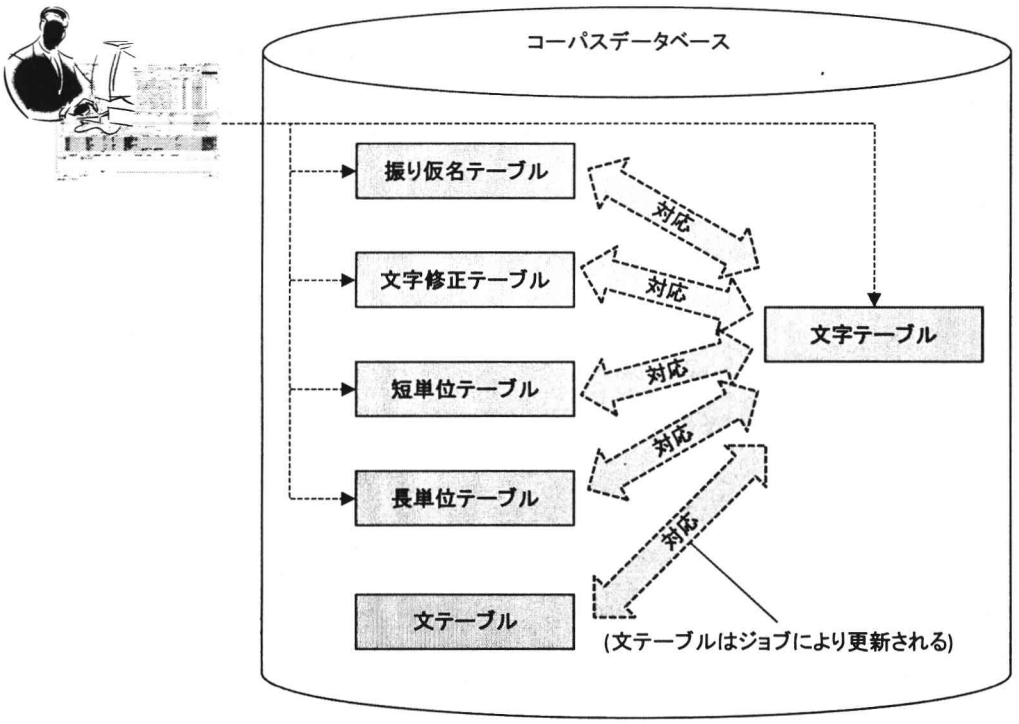


図 64　文字修正時の各テーブルの対応関係

文字修正処理の例として、「にほん」を「にっぽん」に修正する際のテーブル間の対応を示す。

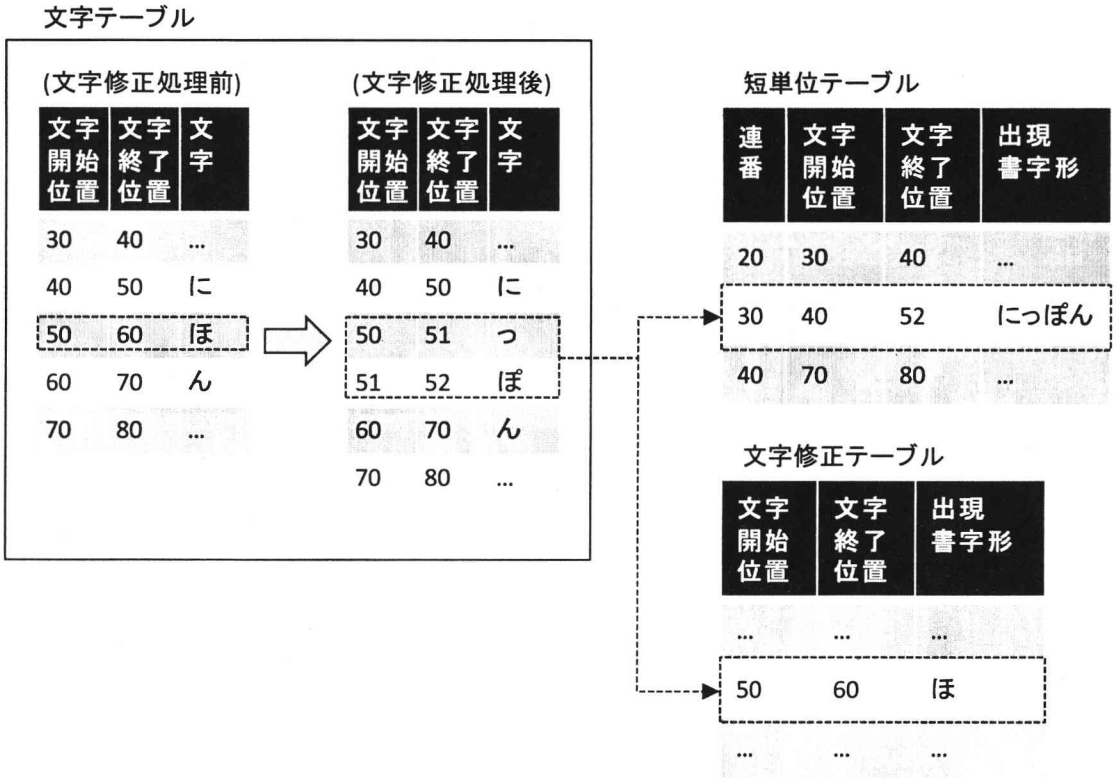


図 65　文字修正処理の例

6.8.長単位モード

6.8.1.長単位モードの概要

「大納言」の長単位モードでは、作業者が短単位との対応を参照しながら、長単位境界の修正と属性の付与を行う。また、更新は文字テーブル等との対応関係が維持されるように処理される。

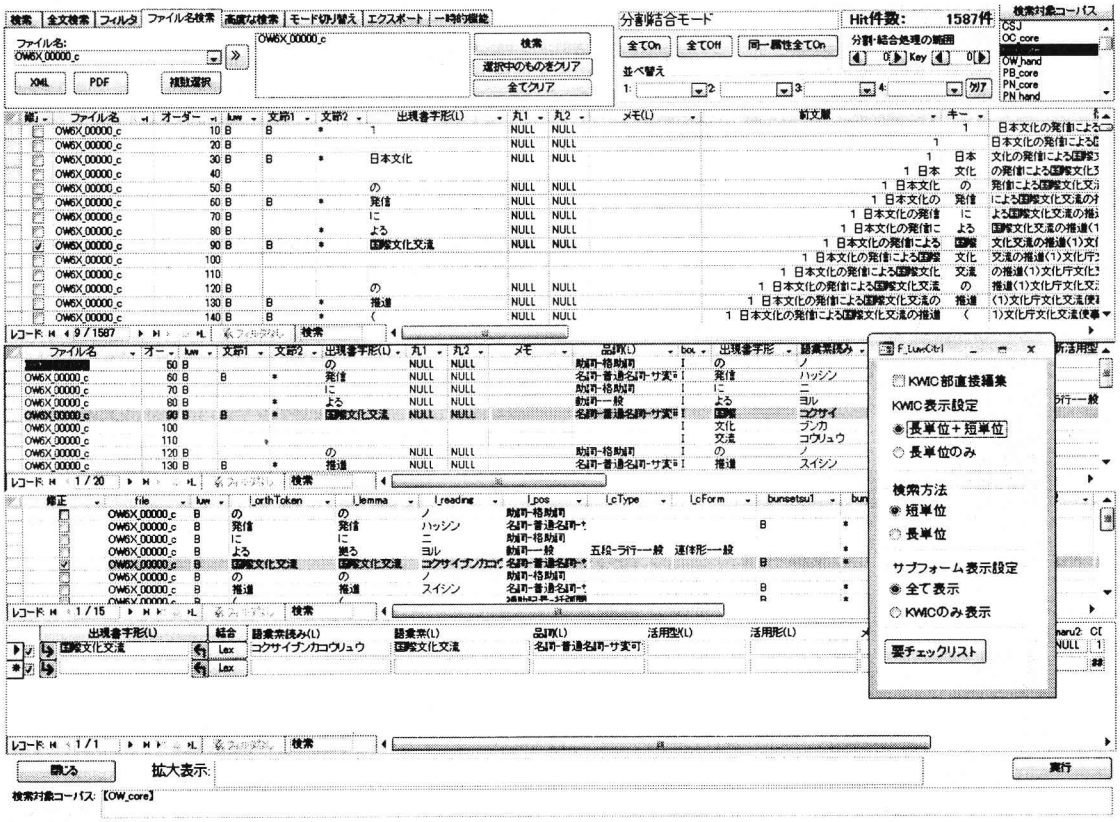


図 66　「大納言」の長単位モード

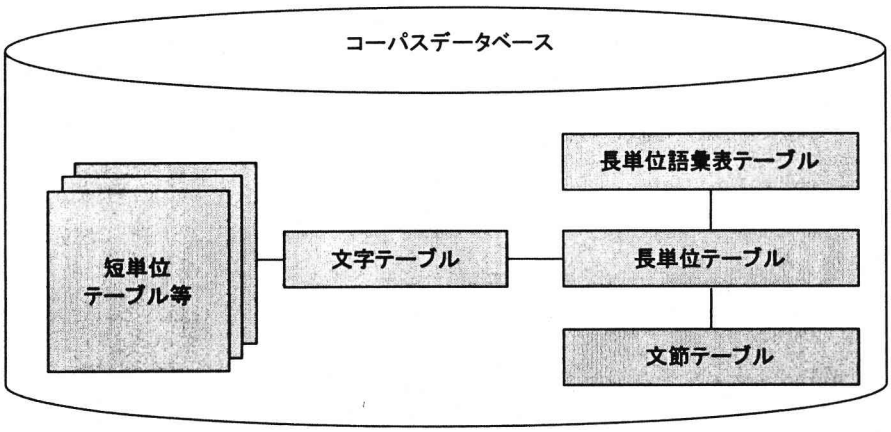


図 67　長単位のテーブル関連図

6.8.2.長単位語彙表について

長単位は短単位をもとにして出現した短単位連続から構成される単位であるが、短単位と同様に、品詞や活用型などの属性を持つ。初期値は長単位解析ツールにより自動で付与されるが、人手による修正を行う必要がある。この際、入力を容易にするために既に出現した長単位については長単位語彙表に格納している。長単位語彙表は属性一意のテーブルであり、作業者はここから選択することにより長単位の属性を付与することができる。
長単位語彙表テーブルの仕様は以下の通りである。長単位のそれぞれの項目の詳細については『形態論情報規程集』を参照のこと。

表 28　長単位語彙表テーブルの項目

項目	説明
ID	連番
長単位出現書字形	短単位の出現形を結合したもの
長単位活用型	末尾の短単位の活用型に概ね一致するが、複合辞など例外あり
長単位活用形	末尾の短単位の活用形に概ね一致するが、複合辞など例外あり
長単位品詞	末尾の短単位の品詞に概ね一致するが、複合辞など例外あり
長単位語彙素読み	活用のない語であれば短単位語彙素読みを結合したものだが、複合動詞などでは再構成する必要がある
長単位語彙素	活用のない語であれば短単位語彙素を結合したものだが、複合動詞などでは再構成する必要がある

なお、長単位語彙表テーブルへのレコードの追加や削除、編集なども大納言上の参照用画面を利用して行う。

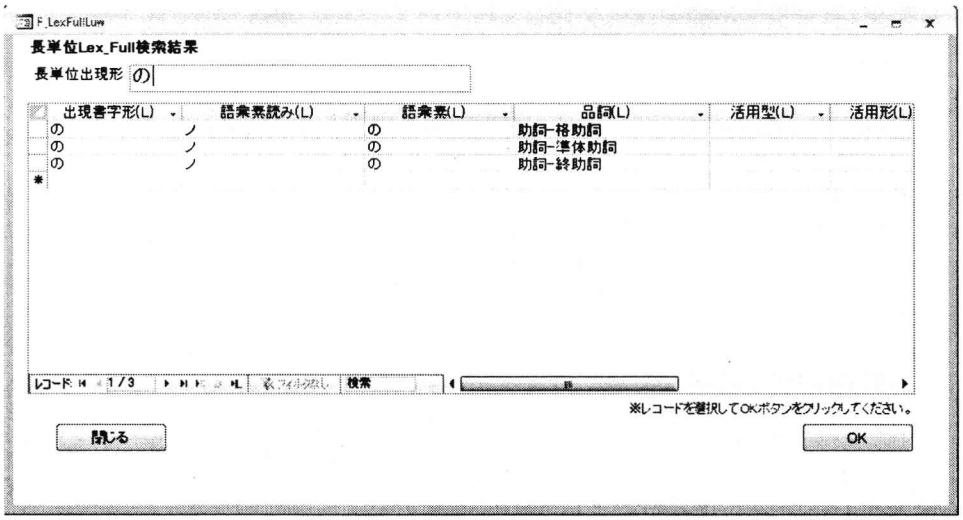


図 68　大納言の長単位語彙表テーブル参照画面

6.8.3.長単位テーブルの更新処理について

長単位の分割結合時には短単位の分割結合時と同様に同一属性一括処理が行える。また、長単位用の文脈チェック処理も行われ、短単位処理と同様に処理前後で文脈が崩れないようにしている。

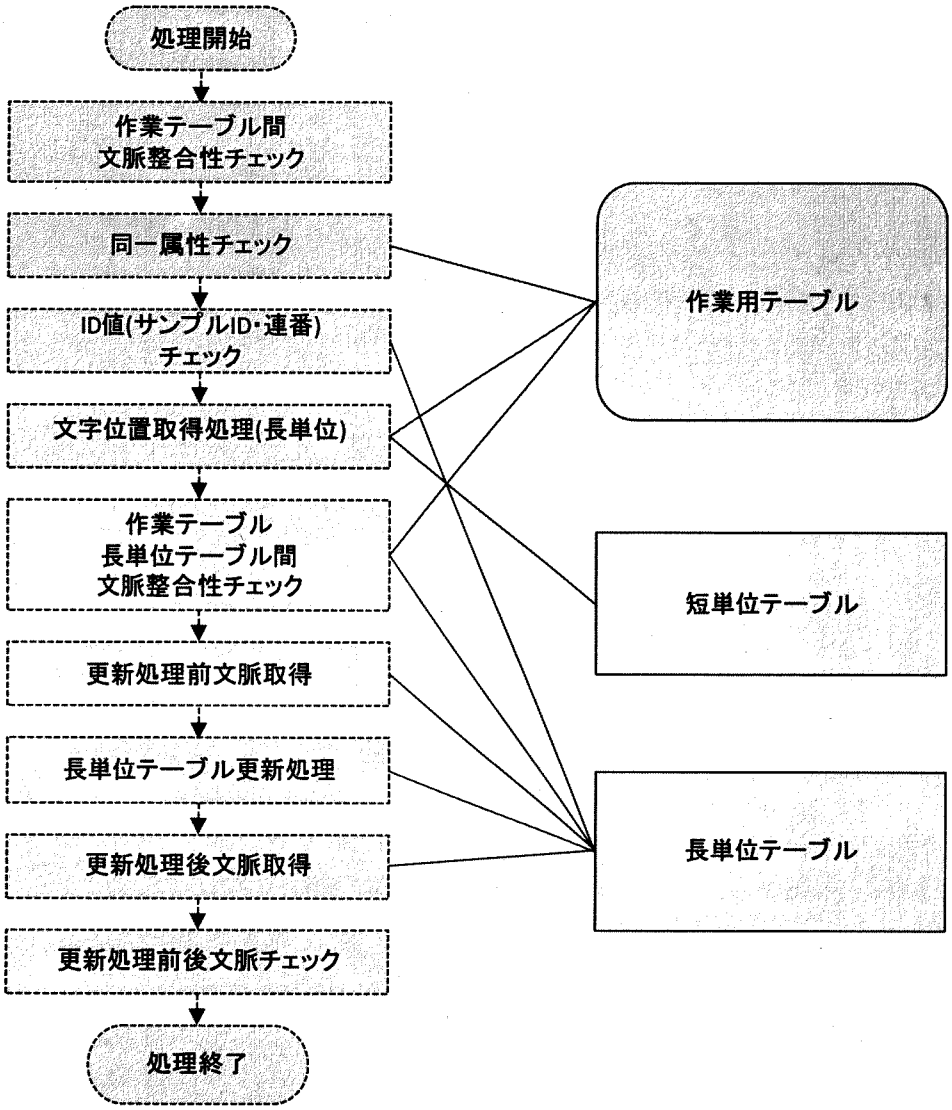


図 69　長単位テーブル更新時の処理の流れ

6.9.文節境界付与モード

文節は、長単位とそれに後続する付属語をひとまとまりにした単位に相当する（文節の定義については『形態論情報規程集』を参照のこと）。「大納言」では、長単位モードを利用して、文節の境界を修正することができる。文節修正をする際は、サンプル ID 検索により語の順番で KWIC を表示させて作業する。

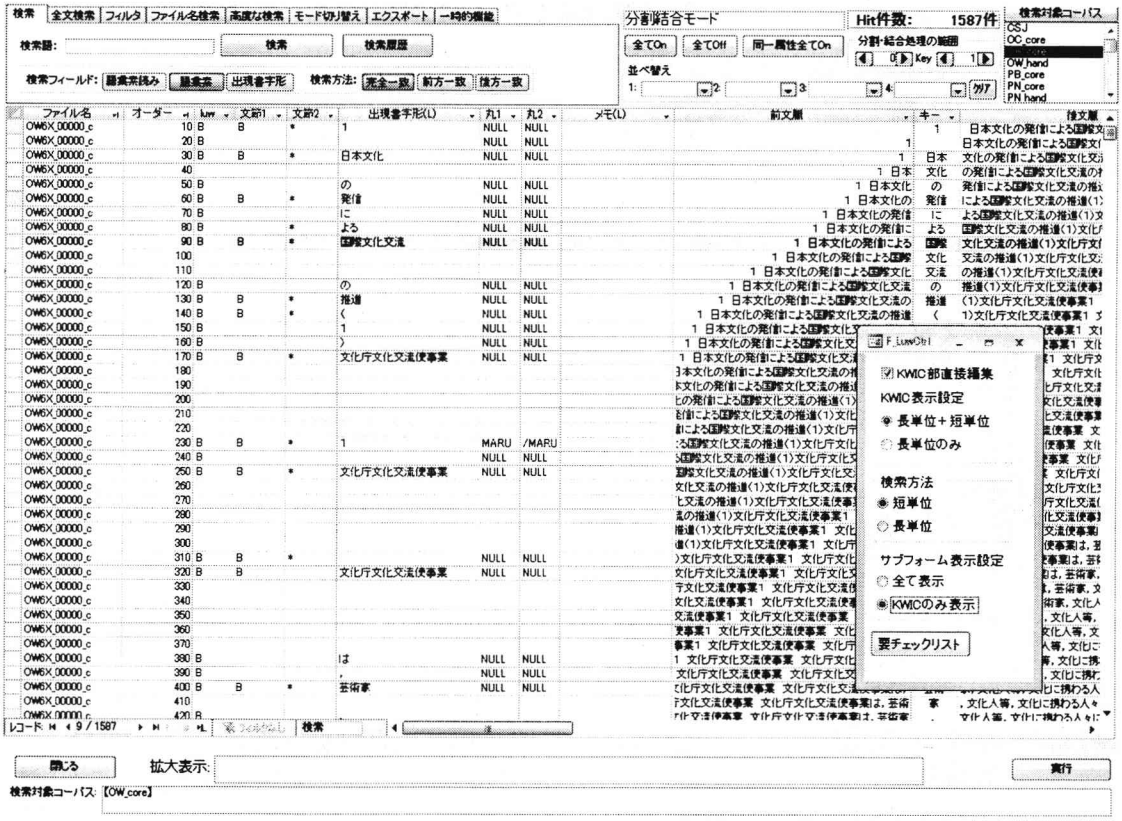


図 70　大納言の文節境界画面

6.10.学習フラグ修正モード

短単位テーブルの「学習フラグ」（形態素解析辞書の学習用コーパスとして利用するかどうかを表す）は、通常の分割結合モードでは修正できない。学習フラグを修正する場合には専用の学習フラグ修正モードを用いる。

この画面では複数レコードを一度に選択し、学習フラグの値を書き込むことができる。書き込みの方法は上書きと追記の二つのモードから選択することができる。

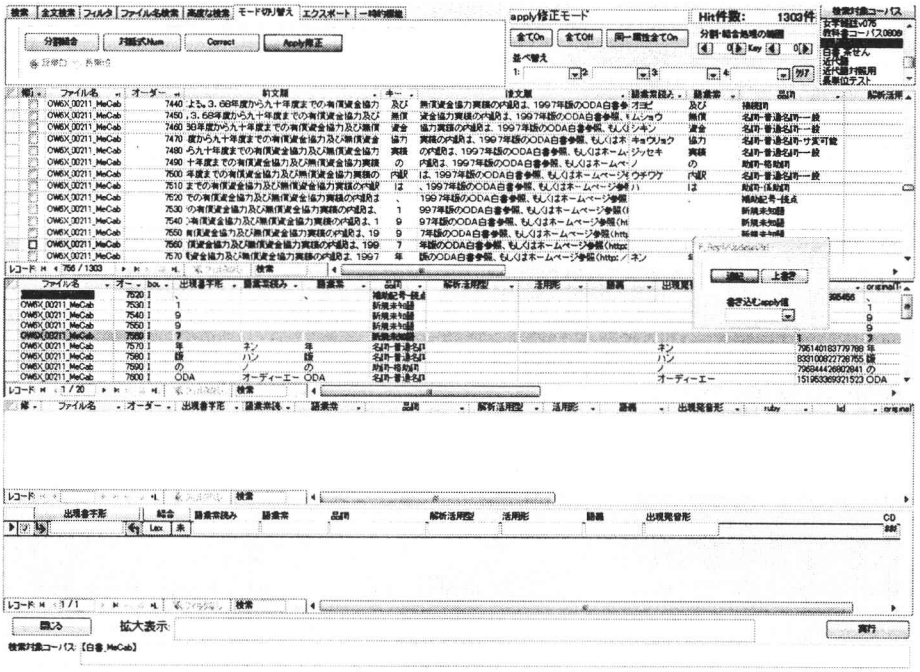


図 71　学習フラグ修正モード画面

7. Web アプリケーション・中納言

7.1. 中納言の概要

コーパス修正ツール・大納言の検索機能は、そのままコーパスを利用した研究に使うことができる。しかし、専用のアプリケーションを必要とするうえ、データベースへの接続に関しても問題を生じる。そこで、Web 上でコーパス検索を可能にする「中納言」を開発している。

中納言は大納言の検索インターフェイスを Web 用に作り直したもので、Web ブラウザ（InternetExplorer・Firefox 等）で利用できる。ブラウザ上での検索や、検索結果のエクスポートが可能となっている。また、中納言は大納言とは別の外部公開用のサーバで稼働しているが、中納言が接続するデータベースの構造は、コーパスデータベースの構造とほぼ同じである。

file	order	mae	key	ato	orthToken	pronToken	pos	reading	sysCType	lemma	cForm	corpusName	start	end	xml	pdf
OW6X_00000_c	14910	で実施しています。 また、	日本	語を第二言語として学ぶ人	日本	ニッポン	名詞-固有名詞-地名-国	ニッポン		日本		OW_core	22720	22740	xml	pdf
PN2c_00017	5480	が役立たなくなると感じ 中国語と	日本	語を学んだ。 すでに日本軍に	日本	ニッポン	名詞-固有名詞-地名-国	ニッポン		日本		PN_core	8500	8520	xml	pdf
PN2c_00017	5750	にした華人虐殺が始まっており、	日本	語を学ぶことが一種の保身のすべ	日本	ニッポン	名詞-固有名詞-地名-国	ニッポン		日本		PN_core	8920	8940	xml	pdf
PN2d_00024_c	3140	れる国立理工科大学校（ポリテクニーク）で	日本	語を教える日本人講師（四十） は	日本	ニッポン	名詞-固有名詞-地名-国	ニッポン		日本		PN_core	5160	5180	xml	pdf
PN3b_00009	3030	5。交差点 心の通った正しい	日本	語を7月二十九日付の	日本	ニッポン	名詞-固有名詞-地名-国	ニッポン		日本		PN_core	4450	4470	xml	pdf
PN3b_00009	4070	になるのも味気ないものです。「正しい	日本	語を使うのはもちろん大事ですが	日本	ニッポン	名詞-固有名詞-地名-国	ニッポン		日本		PN_core	6070	6090	xml	pdf

図 72 「中納言」検索実行画面

「中納言」のシステムは、SQL Server と IIS(Microsoft Internet Information Services), ASP.NET によって実現している。

7.2. 検索機能

中納言には 3 種類の検索方法が用意されており、大納言とほぼ同じ機能が利用できる。

・短単位検索
・全文検索
・高度な検索

また、大納言と同様、検索対象コーパスの指定をすることもできる。

7.3. その他の機能

中納言の検索以外の機能は以下の表の通りである。

表 29　「中納言」の検索以外の機能

機能名	詳細
短単位区切り記号の文脈内表示	前後文脈内に短単位の境界を示す記号を表示することができる。
前後文脈語数指定	前後文脈に表示する語（短単位）数を指定することができる。
固定長・可変長の検索対象指定	検索対象として固定長・可変長・固定長可変長両方を指定することができる。
エクスポート機能	検索結果を Excel 形式でダウンロードできる。

8. ジョブ（定期的自動実行処理）

8.1.ジョブの概要

辞書データベース・コーパスデータベースでは、スケジューリングされた自動実行ジョブによって様々な処理を行っている。基本的には通常行われる作業においてデータベース管理者がデータベースやデータのメンテナンスを行うことはなく、データベースはジョブによって最適な状態が保たれるようになっている。

ジョブによって実行される処理には以下のものがある。

表 30　ジョブによって実行される処理

処理名	処理対象テーブル	実行タイミング
連番の振り直し	短単位テーブル	昼・夜
語種・語形・固定長フラグ・可変長フラグ・語彙素 ID の付与	短単位テーブル	夜
文テーブルのレコード再生成と文開始位置・文終了位置のリセット	文テーブル 短単位テーブル	夜
語彙表の生成	語彙表テーブル	昼・夜
形態素 ID の振り直し	短単位テーブル	夜
属性の振り直し	短単位テーブル	夜
出現頻度の集計	出現頻度表テーブル	夜
書字形構成漢字の再生成	書字形構成漢字テーブル	夜
ログバックアップ処理	-	日中
完全バックアップ	-	毎週
インデックスの再構築	-	毎週

各処理の詳細は以下の通りである。

8.2.連番の振り直し処理

分割結合処理や対話式数字変換処理等をする際に一時的に連番に入力された端数（10 で割り切れない数）を解消する。端数が入力されたサンプルは端数以降の連番がずれることになるため、サンプル単位で処理される。

8.3.見出し語 ID・固定長フラグ・可変長フラグの付与

コーパス内での出現頻度の集計など、データの分析等で頻繁に使われる項目（語種・語形・固定長フラグ・可変長フラグ）については、短単位テーブル内にも格納している。文

字テーブルや語彙表テーブル（辞書データベース）などとデータが重複することになるが、これによってデータ集計時の負荷を大幅に軽減することができる。

また、短単位テーブルの語彙表 ID を専用の ID 変換関数を使用して語彙表 ID に変換することで、短単位テーブルと短単位語彙表テーブルを関連付けすることができるが、ID 変換の負荷が膨大になってしまうため、あらかじめ夜間のジョブ処理によって短単位テーブルに語彙素 ID を格納している。

なお、短単位テーブル上で語彙素 ID を格納している理由は、語の特徴についての情報は辞書データベース上の短単位語彙素テーブルに格納していることが多く、短単位テーブル分析時に短単位語彙素の情報を用いることが多いためである。

8.4. 語彙表の生成

語彙表テーブルは辞書データベース更新時にトリガで自動更新されるが、何らかのトラブル時に語彙表テーブルが正常に更新されない可能性を考慮して、定期的に語彙表テーブルを全件再生成している。実行タイミングは昼/夜としている。

8.5. 属性の振り直し

属性の振り直しは、語彙表テーブルと短単位テーブルにおいて、語彙表 ID が一致しているにもかかわらず品詞等の属性が相違している場合に、語彙表テーブルのデータで短単位テーブルを更新する処理である。この処理によって、語彙表テーブル（辞書データベース）と短単位テーブルの整合性を維持している。

辞書データベースの更新内容はトリガにより即座に語彙表テーブルに反映されるが、処理の負荷を考慮して、リアルタイムで短単位テーブルを更新することはせずに、夜間のジョブ処理によって短単位テーブルと語彙表テーブルの属性値の整合性を維持している。

8.6. 出現頻度の集計

辞書データベースの見出し表修正作業において、短単位テーブルにおける出現頻度を利用することが多いが、やはり出現頻度の集計も負荷のかかる処理であるため、あらかじめ夜間に出現頻度表を生成している。

8.7. 文開始位置リセットと文テーブルのレコード再生成

短単位テーブルと全文検索用の文テーブルは文開始位置・終了位置で関連付けされているため、文テーブルの再生成と文開始位置・終了位置のリセットはセットで行われる。

この処理が必要なのは次のようなサンプルである。まず、インポートした直後のサンプルは全文検索用のデータや、文開始・終了位置がないために処理が必要になる。また、対話式数字変換処理や文字修正処理をした場合については、文（出現書字形）が変更されているので、これについても処理をする必要がある。ただし、この場合は即座に処理せずに、該当する箇所に要再生成のフラグを立てるに止めている。以上のような、文開始・終了位置のないもの、文がないもの、要再生成のフラグが立っているサンプルなどについて、夜間に文開始位置・終了位置のリセットと文テーブルのレコードの再生成処理が行われる。

8.8.ログバックアップ処理

日中は定期的にデータベースのトランザクションログのバックアップ処理が行われる。コーパスデータベース、辞書データベースの両方がトランザクションログバックアップの対象になっている。

8.9.ログの削除・データベースの圧縮・完全バックアップ処理

データベースは徐々に肥大化していきストレージ領域を圧迫してしまうため、定期的にメンテナンスを行う必要がある。特にコーパスデータベースはファイルサイズが非常に巨大であるため、この点は特に重要である。コーパスデータベースでは、毎週末にトランザクションログの削除とデータベースの圧縮、完全バックアップを行うことで、データベースが肥大化しないようにしている。また、作成されたバックアップファイルは物理的に離れた場所にそれぞれ保存され、トラブル時のリスクを分散している。

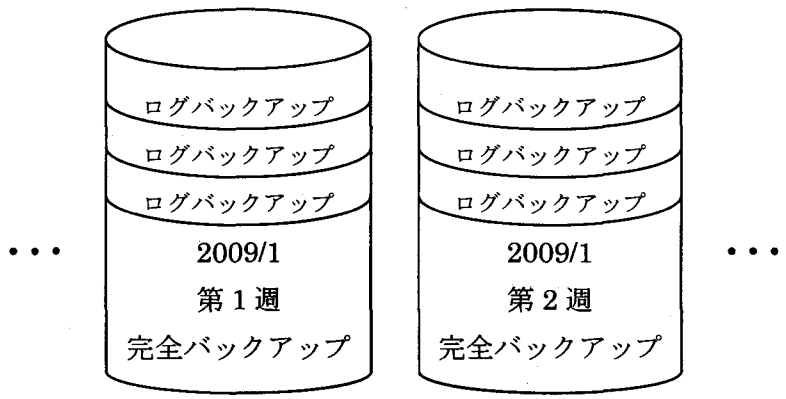


図 73　バックアップ方式の概念図

8.10.インデックスの再構築処理

コーパスデータベースでは検索処理を高速化するためにインデックスを利用しているが、特に短単位テーブルは総レコード数が多いため、インデックスの断片化が起こらないようにすることは重要である。インデックスの再構築処理は非常に時間がかかるため、完全バックアップ同様週末に行っている。またその際には、インデックスのページファイルが分割される頻度を抑えるために、ページファイルに一定の割合で空き領域を設けている。

9. データのインポート・エクスポート

9.1. 概要

ここでは、形態論情報データベース内の各種のデータを取り出したり（エクスポート）、形態素解析結果をデータベースに取り込んだり（インポート）する際の手順と形式について述べる。なかでも重要な次の３つのケースについて説明する。

１．形態素解析辞書の元となるデータ（学習用コーパスと語彙表）のエクスポート
２．XML 形式の BCCWJ サンプルの形態素解析結果のインポート
３．人手修正済みデータ（コアデータ）の XML 形式でのエクスポート

9.2. 形態素解析辞書作成データのエクスポート

形態論情報データベースの役割の一つに、辞書データベースの見出し語と、コーパスデータベースの人手修正データを、形態素解析器（ChaSen, MeCab）の学習用コーパスとして提供することが挙げられる。

現在用いている形態素解析辞書の学習用ツールでは、活用型を展開した語彙表（Lex.txt）と、人手修正コーパス（corpus.txt）を必要とする。いずれもタブ区切りの表形式のテキストで、DBMS の管理ツール（SQL Server Management Studio）上で、SQL 文を実行することよって出力される。形式は次の通りである。なお、いずれのテキストデータも文字符号化方式を UTF-8 に変換する必要がある。

Lex.txt

語彙素読み,語彙素細分類つき語彙素,類,語形（基本形）,出現語形,品詞,活用型,活用形,書字形（基本形）,出現書字形, 発音形（基本形）,出現発音形,語頭変化型,語頭変化形,語頭変化結合型,語末変化型,語末変化形,語末変化結合型,仮名形（基本形）,出現仮名形,アクセント型,アクセント結合型,アクセント修飾型,状態,語種

corpus.txt

コーパス名,サンプル ID,文字開始位置,文字終了位置,文境界,出現書字形,出現発音形,語彙素読み,語彙素細分類つき語彙素,原文文字列,品詞,活用型,活用形,学習フラグ,付加情報,語種

なお、語彙素細分類つき語彙素とは、語彙素細分類の値が空の場合には語彙素を、空でない場合には「語彙素-語彙素細分類」の形式で出力したもの、付加情報は BCCWJ 以外のコーパスで特有の情報を保存するための項目である。

9.3.形態素解析結果のインポート

BCCWJ のサンプルは XML 形式でリリースされる。このデータに形態素解析を施し、形態論情報データベースにインポートする手順について述べる。

形態論情報データベースでは、XML 形式のデータをそのまま取り込むのではなく、関係データベースの表に変換し、それらの表を文字位置をキーにした ID で相互に関係付けることによって、データベース上で XML 文書の構造を再現している。ただし、XML 文書の全てのタグについてではなく、辞書登録やコーパス修正に必要な範囲でのタグについてのみ表として取り込み、それ以外のタグについては元の形のまま保存している（4.1・35 ページ参照）。

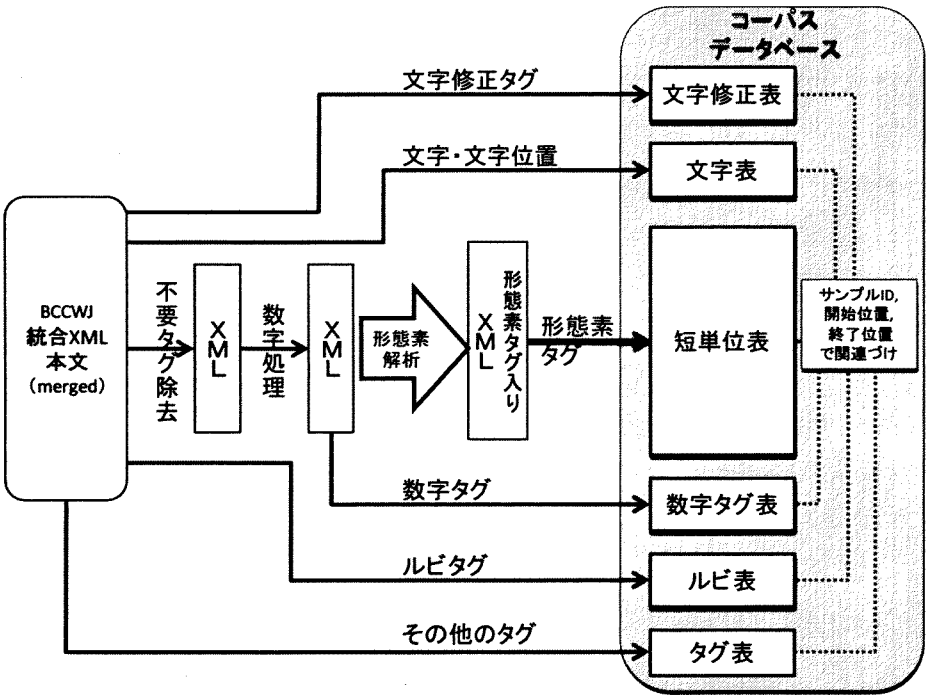


図 74　BCCWJ サンプルの形態素解析とインポート

形態素解析や数字処理の邪魔になるタグの除去や、数字変換などの処理が加わるため、それぞれの表の情報を取り出す段階が異なる（図 74 参照）。タグ・文字テーブルは、元の XML 文書から直接取り出す（したがって、文字テーブルとタグテーブルから XML 文書が完全に再現できる）。数字タグは数字処理後のデータから取り出すことになる。

特に、数字処理では文字がずれる場合があるほか、分子と分母の順番が逆になる場合があるため注意が必要である。

例：１２０円　→　百｜二十｜円

<fraction>１／２</fraction>　→　２｜分｜１

このような文字の変更・移動が起きているため、短単位テーブルは形態素解析結果から単純にとりだすことができない。形態素解析結果を埋め込んだXMLファイルから、原文文字列や数字タグ、分数タグの情報を元に、元の文字との対応を取りながら開始・終了位置を取得する必要がある。この処理は外部のXSLTまたはperlプログラムによって行っている。

図 74の手順で作られた短単位データ、文字データ、文字修正データ、タグデータ、数字データ、振り仮名データを、DBMSの管理ツール（SQL Serer Management Studio）または、大納言のインポート機能（6.3.7・52ページ参照）によってコーパスデータベースに取り込むことで形態素解析結果のインポートが完了する

なお、コーパスデータベースとして、インポートが必須のデータは短単位データと文字データのみである。修正済みデータをXML形式で出力する必要がなければタグなどのデータをインポートする必要はない。

9.4. 人手修正済みデータのエクスポート

取り込んだデータは、人手で修正した後、元のXML文書に形態素タグを埋め込んだXML形式でエクスポートすることができる。DBMSの管理ツール（SQL Server Management Studio）上で、SQL文を実行することよって出力される。

エクスポート用のSQL文では、各テーブルを結合し、データベース内部でXML型のデータとして生成した後、ファイル出力している。データベース内でXML型のデータを生成するため、この時点で整形式のXMLであることが保証される。

テーブルの結合時には、タグテーブルを参照するが、このとき、ルビや数字などの別テーブルで管理されているタグはタグテーブルから出力せず、各テーブルの情報を元にタグを再構成して出力する（サンプルデータ⑲（138ページ）参照）。

資料

① 品詞

短単位語形に入力される「品詞」を示す。詳細は『『現代日本語書き言葉均衡コーパス』形態論情報規程集』を参照。

品詞	大分類	中分類	小分類	細分類	類
名詞-普通名詞-一般	名詞	普通名詞	一般		体
名詞-普通名詞-サ変可能			サ変可能		体
名詞-普通名詞-形状詞可能			形状詞可能		体
名詞-普通名詞-サ変形状詞可能			サ変形状詞可能		体
名詞-普通名詞-副詞可能			副詞可能		体
名詞-普通名詞-連用可能			連用可能		体
名詞-固有名詞-一般		固有名詞	一般		固有名
名詞-固有名詞-人名-一般			人名	一般	人名
名詞-固有名詞-人名-姓			人名	姓	姓
名詞-固有名詞-人名-名			人名	名	名
名詞-固有名詞-地名-一般			地名	一般	地名
名詞-固有名詞-地名-国			地名	国	国
名詞-固有名詞-組織名			組織名		組織名
名詞-数詞		数詞			数
名詞-助動詞語幹		助動詞語幹			体
代名詞	代名詞				体
形状詞-一般	形状詞	一般			相
形状詞-タリ		タリ			相
形状詞-助動詞語幹		助動詞語幹			助動
連体詞	連体詞				相
副詞	副詞				相
接続詞	接続詞				他
感動詞-一般	感動詞	一般			他
感動詞-フィラー		フィラー			他
動詞-一般	動詞	一般			用
動詞-非自立可能		非自立可能			用
形容詞-一般	形容詞	一般			相
形容詞-非自立可能		非自立可能			相
助動詞	助動詞				助動
助詞-格助詞	助詞	格助詞			格助

【資料】

品詞	大分類	中分類	小分類	細分類	類
助詞-副助詞		副助詞			副助
助詞-係助詞		係助詞			係助
助詞-接続助詞		接続助詞			接助
助詞-終助詞		終助詞			終助
助詞-準体助詞		準体助詞			準助
接頭辞	接頭辞				接頭
接尾辞-名詞的-一般	接尾辞	名詞的	一般		接尾体
接尾辞-名詞的-サ変可能			サ変可能		接尾体
接尾辞-名詞的-形状詞可能			形状詞可能		接尾体
接尾辞-名詞的-サ変形状詞可能			サ変形状詞可能		接尾体
接尾辞-名詞的-副詞可能			副詞可能		接尾体
接尾辞-名詞的-助数詞			助数詞		助数
接尾辞-形状詞的		形状詞的			接尾相
接尾辞-動詞的		動詞的			接尾用
接尾辞-形容詞的		形容詞的			接尾相
記号-一般	記号	一般			記号
記号-文字		文字			記号
補助記号-一般	補助記号	一般			補助
空白					補助
補助記号-句点		句点			補助
補助記号-読点		読点			補助
補助記号-括弧開		括弧開			補助
補助記号-括弧閉		括弧閉			補助
補助記号-ＡＡ-一般		ＡＡ	一般		補助
補助記号-ＡＡ-顔文字			顔文字		補助

※ＡＡはアスキーアートの略

② 活用型

以下に形態論情報データベースで用いられる活用型の表を示す（ただし、3.5.7 で述べた特殊な活用型を除く）。表の左側がコーパスデータベースで使われる活用型（＝形態素解析辞書で出力される活用型）、右が辞書データベースに登録する際の活用型（辞書登録活用型）である。

辞書データベースでは、活用型の名前と書字形・発音形を元にして各活用形を展開する必要があるため、単に区別ができればよいコーパスの活用型よりも細かな区別が必要となる。両者の違いは、主に音便形の有無や、その形の違いによるものである。

なお、データベース内部ではこれ以外に、書字形・発音形レベルの差異を反映したさらに詳細な活用型（内部活用型）が用いられることがある（3.5.2 参照）。

動詞（口語）

活用型（コーパス）	辞書登録活用型	補足説明
カ行変格	カ行変格	
サ行変格	サ行変格-スル	一字漢語サ変動詞
	サ行変格-為ル	「する」一語
ザ行変格	サ行変格-ズル	「~ずる」型の一字漢語サ変動詞
上一段-ア行	上一段-ア行	
上一段-カ行	上一段-カ行	
上一段-ガ行	上一段-ガ行	
上一段-ザ行	上一段-ザ行	
上一段-タ行	上一段-タ行	
上一段-ナ行	上一段-ナ行	
上一段-ハ行	上一段-ハ行	
上一段-バ行	上一段-バ行	
上一段-マ行	上一段-マ行	
上一段-ラ行	上一段-ラ行-リル	「足りる」（「足ん-ない」あり）
	上一段-ラ行-一般	
下一段-ア行	下一段-ア行	
下一段-カ行	下一段-カ行	
下一段-ガ行	下一段-ガ行	
下一段-サ行	下一段-サ行-セル	「あわせる」など（連用形「－し」あり）
	下一段-サ行-一般	
下一段-ザ行	下一段-ザ行	
下一段-タ行	下一段-タ行	
下一段-ダ行	下一段-ダ行	
下一段-ナ行	下一段-ナ行	
下一段-ハ行	下一段-ハ行	
下一段-バ行	下一段-バ行	
下一段-マ行	下一段-マ行	
下一段-ラ行-一般	下一段-ラ行-レル	「知れる」など（「知ん-ない」あり）
	下一段-ラ行-一般	
下一段-ラ行-呉レル	下一段-ラ行-呉レル	「呉れる」（命令形が「くれ」）
五段-カ行-イク	五段-カ行-イク	「行く（イク）」（連用形促音便あり）
五段-カ行-ユク	五段-カ行-ユク	「行く（ユク）」（連用形に音便なし）
五段-カ行-一般	五段-カ行-一般	
五段-ガ行	五段-ガ行	
五段-サ行	五段-サ行	
五段-タ行	五段-タ行	
五段-ナ行	五段-ナ行	
五段-バ行	五段-バ行	
五段-マ行	五段-マ行-一般	
	五段-マ行-済ム	「済む」（「すい-ません」）
五段-ラ行-アル	五段-ラ行-アル-サル	「くださる・なさる」（イ音便、命令形「－い」、「－すっ｜た」）
	五段-ラ行-アル-一般	「いらっしゃる・おっしゃる・ござる」（イ音便、命令形「－い」）
五段-ラ行-一般	五段-ラ行-一般	
五段-ワア行-イウ	五段-ワア行-イウ	「言う」（[イ－-マス][ユ－]となる）

【資料】

活用型（コーパス）	辞書登録活用型	補足説明
五段-ワア行-一般	五段-ワア行-一般	
	五段-ワア行-アウ	ウ音便の語形用の区別（以下同じ）
	五段-ワア行-カウ	
	五段-ワア行-ガウ	
	五段-ワア行-タウ	
	五段-ワア行-ダウ	
	五段-ワア行-ツウ	
	五段-ワア行-ナウ	
	五段-ワア行-ハウ	
	五段-ワア行-バウ	
	五段-ワア行-マウ	
	五段-ワア行-ャウ	
	五段-ワア行-ヤウ	
	五段-ワア行-ユウ	
	五段-ワア行-ラウ	
	五段-ワア行-ワウ	

助動詞（口語）

活用型（コーパス）	辞書登録活用型	補足説明
助動詞-ジャ	助動詞-ジャ	
助動詞-タ	助動詞-タ	
助動詞-タイ	助動詞-タイ	
助動詞-ダ	助動詞-ダ	
助動詞-デス	助動詞-デス	
助動詞-ドス	助動詞-ドス	関西（京都）方言
助動詞-ナイ	助動詞-ナイ	
助動詞-ヌ	助動詞-ヌ	
助動詞-ヘン	助動詞-ヘン	関西方言
助動詞-マス	助動詞-マス	
助動詞-ヤ	助動詞-ヤ	
助動詞-ヤス	助動詞-ヤス	「～でやす」
助動詞-ラシイ	助動詞-ラシイ	

形容詞（口語）

活用型（コーパス）	辞書登録活用型	補足説明
形容詞	形容詞-イ段-良イ	「良い（イイ）」（終止連体「ええ」あり）
	形容詞-ウイ	ウ音便の語形用の区別（以下同じ）
	形容詞-オイ	「良い（ヨイ）」
	形容詞-オ段-良イ	「良い（ヨイ）」（終止連体「ええ」あり）
	形容詞-カ	九州方言「きれいか」
	形容詞-カイ	
	形容詞-ガイ	
	形容詞-クイ	
	形容詞-グイ	
	形容詞-コイ	
	形容詞-ゴイ	
	形容詞-サイ	

	形容詞-ザイ	
	形容詞-スイ	
	形容詞-ズイ	
	形容詞-ソイ	
	形容詞-タイ	
	形容詞-ツイ	
	形容詞-トイ	
	形容詞-ドイ	
	形容詞-ナイ	
	形容詞-バイ	
	形容詞-パイ	
	形容詞-ブイ	
	形容詞-ボイ	
	形容詞-ポイ	
	形容詞-マイ	
	形容詞-ムイ	
	形容詞-モイ	
	形容詞-ャイ	
	形容詞-ヤイ	
	形容詞-ユイ	
	形容詞-ョイ	
	形容詞-ヨイ	
	形容詞-ライ	
	形容詞-ルイ	
	形容詞-ロイ	
	形容詞-ワイ	
	形容詞-一般	
	形容詞-無イ	「無い」（終止形「ねえ」あり）

動詞（文語）

活用型（コーパス）	辞書登録活用型	補足説明
文語カ行変格	文語カ行変格	
文語サ行変格	文語サ行変格-ス	
文語ザ行変格	文語サ行変格-ズ	「~ず」型の一字漢語サ変動詞
文語ナ行変格	文語ナ行変格	
文語ラ行変格	文語ラ行変格	
文語上一段-カ行	文語上一段-カ行	
文語上一段-ナ行	文語上一段-ナ行	
文語上一段-マ行	文語上一段-マ行	
文語上一段-ヤ行	文語上一段-ヤ行	
文語上一段-ワ行	文語上一段-ワ行	
文語上二段-タ行	文語上二段-タ行	
文語上二段-ダ行	文語上二段-ダ行	
文語上二段-ハ行	文語上二段-ハ行	
文語上二段-バ行	文語上二段-バ行	
文語上二段-ヤ行	文語上二段-ヤ行	
文語下二段-ア行	文語下二段-ア行	
文語下二段-カ行	文語下二段-カ行	

【資料】

活用型（コーパス）	辞書登録活用型	補足説明
文語下二段-ガ行	文語下二段-ガ行	
文語下二段-サ行	文語下二段-サ行	
文語下二段-ザ行	文語下二段-ザ行	
文語下二段-タ行	文語下二段-タ行	
文語下二段-ダ行	文語下二段-ダ行	
文語下二段-ナ行	文語下二段-ナ行	
文語下二段-ハ行	文語下二段-ハ行-一般	
	文語下二段-ハ行-経	「経（ふ）」
文語下二段-バ行	文語下二段-バ行	
文語下二段-マ行	文語下二段-マ行	
文語下二段-ヤ行	文語下二段-ヤ行	
文語下二段-ラ行	文語下二段-ラ行	
文語四段-カ行	文語四段-カ行	
文語四段-ガ行	文語四段-ガ行	
文語四段-サ行	文語四段-サ行	
文語四段-タ行	文語四段-タ行	
文語四段-ハ行	文語四段-ハ行-アウ	
	文語四段-ハ行-イウ	
	文語四段-ハ行-カウ	
	文語四段-ハ行-ガウ	
	文語四段-ハ行-タウ	
	文語四段-ハ行-ダウ	
	文語四段-ハ行-ナウ	
	文語四段-ハ行-ハウ	
	文語四段-ハ行-バウ	
	文語四段-ハ行-マウ	
	文語四段-ハ行-ヤウ	
	文語四段-ハ行-ラウ	
	文語四段-ハ行-ワウ	
	文語四段-ハ行-一般	
	文語四段-ハ行-チョウ	「てふ」（「といふ」の融合形）
文語四段-バ行	文語四段-バ行	
文語四段-マ行	文語四段-マ行	
文語四段-ラ行	文語四段-ラ行	

助動詞（文語）

活用型（コーパス）	辞書登録活用型	補足説明
文語助動詞-キ	文語助動詞-キ	
文語助動詞-ケム	文語助動詞-ケム	
文語助動詞-ケリ	文語助動詞-ケリ	
文語助動詞-ゴトシ	文語助動詞-ゴトシ	
文語助動詞-ザマス	文語助動詞-ザマス	
文語助動詞-ザンス	文語助動詞-ザンス	
文語助動詞-ズ	文語助動詞-ズ	
文語助動詞-タリ-完了	文語助動詞-タリ-完了	
文語助動詞-タリ-断定	文語助動詞-タリ-断定	
文語助動詞-ツ	文語助動詞-ツ	

文語助動詞-ナリ-伝聞	文語助動詞-ナリ-伝聞	
文語助動詞-ナリ-断定	文語助動詞-ナリ-断定	
文語助動詞-ヌ	文語助動詞-ヌ	
文語助動詞-ベシ	文語助動詞-ベシ	
文語助動詞-マジ	文語助動詞-マジ	
文語助動詞-ム	文語助動詞-ム	
文語助動詞-ラシ	文語助動詞-ラシ	
文語助動詞-ラム	文語助動詞-ラム	
文語助動詞-リ	文語助動詞-リ	
文語助動詞-ンス	文語助動詞-ンス	近世上方語
無変化型	無変化型	

形容詞（文語）

活用型（コーパス）	辞書登録活用型	補足説明
文語形容詞-ク	文語形容詞-ク	
文語形容詞-シク	文語形容詞-シク	
文語形容詞-ジク	文語形容詞-ジク	「いみじ」など
文語形容詞-多シ	文語形容詞-多シ	「多し」（終止「多かり」）

③ 活用形

以下に形態論情報データベースで用いられる活用形の表を示す。活用形は自動で展開されるため、辞書登録ユーザが直接入力することはない。

大分類	活用形	補足説明
語幹	語幹-サ	形容詞「無い」「良い」に、様態の助動詞「そうだ」が接続するときの形（「無さ-そうだ」「良さ-そうだ」）
	語幹-一般	
未然形	未然形-サ	サ変（ザ変）に、助動詞「せる」「れる」が接続するときの形（「さ-せる」「さ-れる」）
	未然形-セ	サ変（ザ変）に、助動詞「ず」が接続するときの形（せ-ず）
	未然形-一般	
	未然形-撥音便	ラ行五段活用動詞の一部で起こる撥音便（「知ん-ない」）
	未然形-補助	形容詞カリ活用未然形（「少なから-ず」）
意志推量形	意志推量形	意志・推量の助動詞「う」「よう」が接続した形全体（「行こう」「見よう」）
連用形	連用形-イ音便	
	連用形-ウ音便	
	連用形-キ接続	文語サ行四段活用の動詞に過去の文語助動詞「き」が接続する場合に「～せ」の形を取ることがある（「許せ-し」）cf.「文法上許容スベキ事項」
	連用形-ト	断定の文語助動詞「たり」の連用形「と」

【資料】

大分類	活用形	補足説明
	連用形-ニ	断定の助動詞「だ」・文語助動詞「なり」の連用形「に」
	連用形-一般	
	連用形-促音便	
	連用形-撥音便	
	連用形-省略	関西方言などで形容詞連用形が省略された形をとることがある（「欲し-ない」）
	連用形-融合	断定の助動詞「だ」の連用形に後続する係助詞「は」が融合した形（「じゃ」）
	連用形-補助	文語形容詞・文語助動詞「ず」のカリ活用連用形（「無かり」「ざり」）
終止形	終止形-ウ音便	文語ハ行四段活用動詞の終止形がウ音便化することがある（「給う［タモー］」「候［ソーロー］」）
	終止形-一般	
	終止形-促音便	形容詞の「高っ」「痛っ」などの形
	終止形-撥音便	助動詞「ず」の終止形に撥音便形がある（「（しませ）ん」）また関西方言などで撥音便形になることがある（「てん（な）」）
	終止形-融合	断定の助動詞「だ」の終止形に前接する「と」の音と融合した形（「（何のこっ）ちゃ」）
	終止形-補助	文語形容詞「多し」のカリ活用終止形（「多かり」）
連体形	連体形-ウ音便	文語ハ行四段活用動詞の連体形がウ音便化することがある（「給う［タモー］」「候［ソーロー］」）
	連体形-一般	
	連体形-撥音便	助動詞「ず」の連体形がしばしば「ん」となるほか、動詞でも「すん（の）」のように準体助詞「の」の前で撥音になるまた文語助動詞「む」「けむ」の連体形が「ん」となる
	連体形-省略	
	連体形-補助	
已然形	已然形	
	已然形-一般	
	已然形-補助	
仮定形	仮定形-一般	
	仮定形-融合	
命令形	命令形	
	命令形-一般	
ク語法	ク語法	文語専用

④ 語頭変化表

語頭変化型	語頭変化形	語頭変化形 subID	語頭語形	代表性
カ濁	基本形	1	カ	True
	濁音形	2	ガ	False
キ濁	基本形	1	キ	True
	濁音形	2	ギ	False
ク濁	基本形	1	ク	True
	濁音形	2	グ	False
ケ濁	基本形	1	ケ	True
	濁音形	2	ゲ	False
コ濁	基本形	1	コ	True
	濁音形	2	ゴ	False
サ濁	基本形	1	サ	True
	濁音形	2	ザ	False
シ濁	基本形	1	シ	True
	濁音形	2	ジ	False
ス濁	基本形	1	ス	True
	濁音形	2	ズ	False
セ濁	基本形	1	セ	True
	濁音形	2	ゼ	False
ソ濁	基本形	1	ソ	True
	濁音形	2	ゾ	False
タ濁	基本形	1	タ	True
	濁音形	2	ダ	False
チ濁	基本形	1	チ	True
	濁音形	2	ヂ	False
ツ濁	基本形	1	ツ	True
	濁音形	2	ヅ	False
テ濁	基本形	1	テ	True
	濁音形	2	デ	False
ト濁	基本形	1	ト	True
	濁音形	2	ド	False
ハ半濁	半濁音形	3	パ	False
	基本形	1	ハ	True
ハ混合	半濁音形	3	パ	False
	基本形	1	ハ	True
	濁音形	2	バ	False
ハ濁	基本形	1	ハ	True
	濁音形	2	バ	False
ヒ半濁	半濁音形	3	ピ	False
	基本形	1	ヒ	True
ヒ混合	半濁音形	3	ピ	False
	基本形	1	ヒ	True
	濁音形	2	ビ	False
ヒ濁	基本形	1	ヒ	True

【資料】

語頭変化型	語頭変化形	語頭変化形 subID	語頭語形	代表性
	濁音形	2	ビ	False
フ半濁	半濁音形	3	プ	False
	基本形	1	フ	True
フ混合	半濁音形	3	プ	False
	基本形	1	フ	True
	濁音形	2	ブ	False
フ濁	基本形	1	フ	True
	濁音形	2	ブ	False
ヘ半濁	半濁音形	3	ペ	False
	基本形	1	ヘ	True
ヘ混合	半濁音形	3	ペ	False
	基本形	1	ヘ	True
	濁音形	2	ベ	False
ヘ濁	基本形	1	ヘ	True
	濁音形	2	ベ	False
ホ半濁	半濁音形	3	ポ	False
	基本形	1	ホ	True
ホ混合	半濁音形	3	ポ	False
	基本形	1	ホ	True
	濁音形	2	ボ	False
ホ濁	基本形	1	ホ	True
	濁音形	2	ボ	False
ワ混合	半濁音形	3	パ	False
	基本形	1	ワ	True
	濁音形	2	バ	False

⑤ 語末変化表

語末変化型	語末変化形	語末変化形 subID	語末語形	代表性	語末発音形
ア長促添	基本形	1		True	
	長音添加形	4	ア	False	ー
	促音添加形	5	ッ	False	ッ
ア長促撥添	基本形	1		True	
	長音添加形	4	ア	False	ー
	促音添加形	5	ッ	False	ッ
	撥音添加形	6	ン	False	ン
ア長添	基本形	1		True	
	長音添加形	4	ア	False	ー
イ長促添	基本形	1		True	
	長音添加形	4	イ	False	ー
	促音添加形	5	ッ	False	ッ
イ長促撥添	基本形	1		True	
	長音添加形	4	イ	False	ー

語末変化型	語末変化形	語末変化形 subID	語末語形	代表性	語末発音形
	促音添加形	5	ッ	False	ッ
	撥音添加形	6	ン	False	ン
イ長添	基本形	1		True	
	長音添加形	4	イ	False	ー
ウ長促添	基本形	1		True	
	長音添加形	4	ウ	False	ー
	促音添加形	5	ッ	False	ッ
ウ長促撥添	基本形	1		True	
	長音添加形	4	ウ	False	ー
	促音添加形	5	ッ	False	ッ
	撥音添加形	6	ン	False	ン
ウ長添	基本形	1		True	
	長音添加形	4	ウ	False	ー
エ長促添	基本形	1		True	
	長音添加形	4	エ	False	ー
	促音添加形	5	ッ	False	ッ
エ長促撥添	基本形	1		True	
	長音添加形	4	エ	False	ー
	促音添加形	5	ッ	False	ッ
	撥音添加形	6	ン	False	ン
エ長添	基本形	1		True	
	長音添加形	4	エ	False	ー
オ長促添	基本形	1		True	
	長音添加形	4	オ	False	ー
	促音添加形	5	ッ	False	ッ
オ長促撥添	基本形	1		True	
	長音添加形	4	オ	False	ー
	促音添加形	5	ッ	False	ッ
	撥音添加形	6	ン	False	ン
オ長添	基本形	1		True	
	長音添加形	4	オ	False	ー
キ促	基本形	1	キ	True	キ
	促音形	2	ッ	False	ッ
ク促	基本形	1	ク	True	ク
	促音形	2	ッ	False	ッ
チ促	基本形	1	チ	True	チ
	促音形	2	ッ	False	ッ
ツ促	基本形	1	ツ	True	ツ
	促音形	2	ッ	False	ッ
十促	基本形	1	ュウ	True	ュー
	促音形	2	ッ	False	ッ
	促音形	3	ュッ	False	ュッ
促添	基本形	1		True	
	促音添加形	5	ッ	False	ッ
促撥添	基本形	1		True	
	促音添加形	5	ッ	False	ッ

語末変化型	語末変化形	語末変化形 subID	語末語形	代表性	語末発音形
	撥音添加形	6	ン	False	ン

⑥ 見出し語の出典

短単位見出し語テーブルに共通で付与される属性のうち、記号で表される「出典」の値の一覧を示す（主なもののみ）。

値	出典
c	CSJ
b	BCCWJ 書籍
w	BCCWJ 白書
n	BCCWJ 新聞
y	BCCWJ Web データ
近	近代語データ
太	太陽コーパス

⑦ 見出し語の状態

短単位見出し語テーブルに共通で付与される属性のうち、記号で表される「状態」の値の一覧を示す。

値	見出し語の状態
仮	確認が終わるまで形態素解析辞書には出力しない（仮登録）
Z	コアデータに出現したため登録しているが、解析辞書には出力しない
y	BCCWJ のサンプル解析でのみ利用し、一般用の解析辞書には出力しない
k	近代語用の解析辞書にのみ出力する
c	近代語用の解析辞書には出力しない

※k, c は「近代文語 UniDic」用の値

⑧ オリジナル関数一覧

コーパスデータベース

関数名	引数	説明
前文脈生成関数	サンプル ID・連番	KWIC の前文脈を返す関数。
後文脈生成関数	サンプル ID・連番	KWIC の後文脈を返す関数。
検索語文中出現数カウント関数	文・検索語	全文検索時に使われる関数。文中の検索語出現数をカウントする。
ID 変換関数	変換前項目名・変換後項目名・ID	語彙素 ID・語形 ID・書字形 ID・発音形 ID・語彙表 ID を他の ID に変換する関数。
文字修正情報取得関数	文字開始位置・文字終了位置・サンプル ID	文字修正テーブルから該当箇所の文字修正情報を取得する関数。
数字情報取得関数	文字開始位置・文字終了位置・サンプル ID	数字テーブルから該当箇所の数字情報を取得する関数。
振り仮名情報取得関数	文字開始位置・文字終了位置・サンプル ID	振り仮名テーブルから該当箇所の振り仮名情報を取得する関数。
活用型変換関数	書字形・発音形・活用型	辞書データベースの活用型から語彙表を作成するのに必要な解析活用型に変換する関数。
活用型書字形変換関数	活用型・比較する活用型・段	活用型に付与する詳細情報を生成する関数。活用型とこの詳細情報から解析活用型が生成される。
カタカナひらがな変換関数	文字列	文字列内のカタカナを平仮名に変換する関数。
語頭語末変化関数	語頭変化型・語頭変化形 ID・語末変化型・語末変化形 ID・文字列・変化レベル	文字列を語頭語末変化させて返す関数。

【資料】

辞書データベース

関数名	引数	説明
ひらがなカタカナ変換関数	文字列	文字列内の平仮名をカタカナに変換する関数。
アクセント結合型取得関数	文字列・アクセント型	文字列のアクセント結合型を取得する関数。
モーラ数取得関数	文字列	文字列内のモーラ数を取得する関数。
アルファベット全角半角変換関数	文字列	文字列内の半角アルファベットを全角アルファベットに変換する関数。
出現頻度カウント関数	コーパス名・開始語彙表 ID, 終了語彙表 ID・固定長可変長	短単位テーブルにおける出現頻度をカウントする関数。
語頭濁音形変換関数	文字列	文字列の語頭にあるカタカナの濁音を清音に変換する関数。

⑨ ストアドプロシージャー一覧

辞書データベース

ストアドプロシージャ名	引数	説明
書字形構成漢字 ストアドプロシージャ	モード	書字形から漢字を抽出して音訓等種別と音訓を付与して書字形構成漢字テーブルに格納するストアドプロシージャ。
漢字頻度集計 ストアドプロシージャ	固定長・可変長・集計条件	漢字音訓頻度表生成の第一段階。漢字・音訓等種別・音訓の出現頻度を集計するストアドプロシージャ。
漢字頻度書式修正 ストアドプロシージャ	なし	漢字音訓頻度表生成の第二段階。漢字頻度集計結果を利用して漢字音訓頻度表用の表記を生成するストアドプロシージャ。
短単位出現頻度集計 ストアドプロシージャ	レベル（語彙素・語形・書字形）	コーパス内の語の出現頻度表を生成するストアドプロシージャ。

コーパスデータベース

ストアドプロシージャ名	引数	説明
学習フラグ更新 ストアドプロシージャ	モード・追記文字・ユーザ名	大納言の学習フラグ修正モードで使用され、短単位テーブルの状態フラグを更新するストアドプロシージャ。
短単位分割結合 ストアドプロシージャ	サンプル ID・連番・開始処理範囲・終了処理範囲・ユーザ名・モード	大納言の短単位モードにて使用される。短単位の分割結合・文字修正・対話式数字変換処理を行う。内部で文字位置割振りストアドプロシージャを呼び出している。
DB バックアップ ストアドプロシージャ	モード・データベース名・バックアップ先1・バックアップ先2・バックアップ先3	データベースのバックアップ処理をするストアドプロシージャ。中小サイズ用。
DB バックアップ ストアドプロシージャ （巨大 DB 用）	モード・データベース名・バックアップ先1・バックアップ先2・バックアップ先3	データベースのバックアップ処理をするストアドプロシージャ。巨大サイズ用。
DB メンテナンス ストアドプロシージャ	データベース名・バックアップ先1・バックアップ先2・バックアップ先3	データベースのログの削除・圧縮・バックアップをするストアドプロシージャ。中小サイズ用。
DB メンテナンス ストアドプロシージャ （巨大 DB 用）	データベース名・バックアップ先1・バックアップ先2・バックアップ先3	データベースのログの削除・圧縮・バックアップをするストアドプロシージャ。巨大サイズ用）。

【資料】

ストアドプロシージャ名	引数	説明
インデックス再構築 ストアドプロシージャ	なし	データベース内の全てのテーブルのインデックスを再構築するストアドプロシージャ。
データ削除 ストアドプロシージャ	削除単位・削除対象	データを削除するストアドプロシージャ。
語彙表不整合抽出 ストアドプロシージャ	なし	語彙表テーブルと短単位テーブルの不整合を抽出するストアドプロシージャ。
データ取り込み ストアドプロシージャ	コーパス名・ユーザ名	テキストファイルをインポートしてコーパスの各テーブルに格納するストアドプロシージャ。
高度な検索 ストアドプロシージャ	検索語・ユーザ名・モード	短単位の高度な検索を行い、結果を作業テーブルに格納するストアドプロシージャ。
短単位検索 ストアドプロシージャ	検索語・検索タイプ・検索フィールド・ユーザ名・コーパス名	短単位の検索を行い、結果を作業テーブルに格納するストアドプロシージャ。
全文検索 ストアドプロシージャ	検索語・検索対象コーパス名・ユーザ名	文テーブルに対して全文検索を行い、結果を作業テーブルに格納するストアドプロシージャ。
語彙表生成 ストアドプロシージャ	更新レベル・削除するID・挿入するID	語彙表を生成するストアドプロシージャ。特定の ID の範囲のみの再生成と全件再生成を行える。
短単位文字位置割振り ストアドプロシージャ	サンプル ID・キーオーダー・ユーザ名・前語数・後語数・モード	短単位分割結合ストアドプロシージャで呼び出されるストアドプロシージャ。文字開始位置・終了位置を作業テーブルに入力する。短単位分割結合用。
文字修正箇所の原文文字列取得ストアドプロシージャ	サンプル ID・キーオーダー・ユーザ名・前語数・後語数	文字修正処理された箇所のオリジナルの文字列を取得するストアドプロシージャ(文字修正処理モード用)。
数字変換箇所の原文文字列取得ストアドプロシージャ	サンプル ID・キーオーダー・ユーザ名・前語数・後語数	数字変換処理された箇所のオリジナルの文字列を取得するストアドプロシージャ（対話式数字変換処理モード用）。
連番振り直し ストアドプロシージャ	サンプル ID	短単位テーブルの連番を振り直すストアドプロシージャ。
属性更新 ストアドプロシージャ	なし	語彙表テーブルと短単位テーブルの齟齬を解消するストアドプロシージャ。

ストアドプロシージャ名	引数	説明
周辺語取得 ストアドプロシージャ	連番・サンプル ID・ユーザ名・一時テーブル接尾辞	指定した語の周辺の語を短単位テーブルから取得するストアドプロシージャ。
作業テーブル間データコピー ストアドプロシージャ	ユーザ名	一時テーブル(周辺語)を一時テーブル(誤)にコピーするストアドプロシージャ。
短単位作業テーブル(KWIC 用)生成ストアドプロシージャ	ユーザ名	KWIC を格納する作業テーブルを生成するストアドプロシージャ。
短単位作業テーブル生成 ストアドプロシージャ	ユーザ名・接尾辞	作業テーブル(誤)と作業テーブル(正)を生成するストアドプロシージャ。
長単位取得 ストアドプロシージャ	ユーザ名	作業テーブルに格納された短単位に対応する長単位レコードを生成するストアドプロシージャ。
長単位周辺語取得 ストアドプロシージャ	ユーザ名	大納言で選択中の長単位の周辺の長単位を取得するストアドプロシージャ。
全文検索用データ整備 ストアドプロシージャ	モード	全文検索で使用されるデータを整えるストアドプロシージャ。短単位テーブルの文開始・終了位置と文テーブルを更新する。
長単位文字位置割振り ストアドプロシージャ	ユーザ名	長単位の文字開始位置・終了位置を作業テーブルに入力するストアドプロシージャ。
文節更新 ストアドプロシージャ	ユーザ名	文節テーブルを更新するストアドプロシージャ。
長単位更新 ストアドプロシージャ	ユーザ名	長単位テーブルを更新するストアドプロシージャ。

⑩ テーブル一覧

辞書データベース

テーブル名		短単位語彙素テーブル	
説明		3.2.2 短単位語彙素テーブル(10 ページ)参照	
No	フィールド名	データ型	説明
1	語彙素 ID	int identity	
2	語彙素	Nvarchar	
3	語彙素読み	Nvarchar	
4	類	Nvarchar	
5	出典	Nvarchar	
6	状態	Nvarchar	
7	コメント	Ntext	
8	評価	Nvarchar	
9	原語表記	Nvarchar	
10	語彙素細分類	nvarchar	
11	語種	nvarchar	
12	更新作業者	nvarchar	
13	更新日時	datetime	
14	最小単位	nvarchar	
15	最小単位数	int	

テーブル名		短単位語形テーブル	
説明		3.2.3 短単位語形テーブル(12 ページ)参照	
No	フィールド名	データ型	説明
1	語形 ID	int	
2	語彙素 ID	int	
3	語形 SubID	int	
4	語形	nvarchar	
5	品詞	nvarchar	
6	活用型	nvarchar	
7	語頭変化型	nvarchar	
8	語頭変化結合型	nvarchar	
9	語末変化型	nvarchar	
10	語末変化結合型	nvarchar	
11	代表性	bit	
12	出典	nvarchar	
13	状態	nvarchar	
14	コメント	ntext	
15	評価	nvarchar	
16	更新作業者	nvarchar	
17	更新日時	datetime	

テーブル名		短単位書字形テーブル	
説明		3.2.4 短単位書字形テーブル(14 ページ)参照	
No	フィールド名	データ型	説明
1	書字形 ID	bigint	
2	語形 ID	int	
3	書字形 SubID	int	
4	書字形	nvarchar	
5	活用型書字形	nvarchar	
6	仮名形	nvarchar	
7	代表性	bit	
8	出典	nvarchar	
9	状態	nvarchar	
10	コメント	ntext	
11	評価	nvarchar	
12	更新作業者	nvarchar	
13	更新日時	datetime	

テーブル名		短単位発音形テーブル	
説明		3.2.5 短単位発音形テーブル(15 ページ)参照	
No	フィールド名	データ型	説明
1	発音形 ID	bigint	
2	語形 ID	int	
3	発音形 SubID	int	
4	発音形	nvarchar	
5	活用型発音形	nvarchar	
6	アクセント型	nvarchar	
7	アクセント結合型	nvarchar	
8	代表性	bit	
9	出典	nvarchar	
10	アクセント型出典	nvarchar	
11	状態	nvarchar	
12	コメント	ntext	
13	評価	nvarchar	
14	更新作業者	nvarchar	
15	更新日時	datetime	

テーブル名		語頭変化表テーブル	
説明		3.4.2 語頭変化(19 ページ)参照	
No	フィールド名	データ型	説明
1	語頭変化型	nvarchar	
2	語頭変化形	nvarchar	
3	語頭変化形 subID	tinyint	
4	語頭語形	nvarchar	
5	代表性	bit	

【資料】

テーブル名		語末変化表テーブル	
説明		3.4.3 語末変化(19 ページ)参照	
No	フィールド名	データ型	説明
1	語末変化型	nvarchar	
2	語末変化形	nvarchar	
3	語末変化形 subID	tinyint	
4	語末語形	nvarchar	
5	代表性	bit	
6	語末発音形	nvarchar	

テーブル名		活用表テーブル	
説明		3.5.4 活用表(23 ページ)参照	
No	フィールド名	データ型	説明
1	活用型	nvarchar	
2	活用形	nvarchar	
3	活用語尾	nvarchar	
4	代表性	bit	
5	活用語尾書字形	nvarchar	
6	活用語尾発音形	nvarchar	
7	活用語尾仮名形	nvarchar	
8	アクセント修飾型	nvarchar	
9	詳細活用形	nvarchar	
10	状態	nvarchar	

テーブル名		活用型テーブル	
説明		3.5.6 活用形テーブルと活用型テーブル(24 ページ)参照	
No	フィールド名	データ型	説明
1	活用型	nvarchar	
2	大分類	nvarchar	
3	行分類	nvarchar	
4	段分類	nvarchar	
5	小分類	nvarchar	

テーブル名		活用形テーブル	
説明		3.5.6 活用形テーブルと活用型テーブル(24 ページ)参照	
No	フィールド名	データ型	説明
1	活用形 ID	int	
2	詳細活用形	nvarchar	
3	大分類	nvarchar	
4	小分類	nvarchar	
5	活用形	nvarchar	

テーブル名		活用型簡略化テーブル	
説明		3.5.3 活用型簡略化(22 ページ)参照	
No	フィールド名	データ型	説明
1	辞書登録活用型	nvarchar	
2	内部活用型	nvarchar	
3	活用型	nvarchar	コーパス活用型

テーブル名		品詞テーブル	
説明		品詞を入力するための参照用データ	
No	フィールド名	データ型	説明
1	品詞 ID	int	主キー
2	品詞	nvarchar	品詞全体
3	大分類	nvarchar	品詞の第 1 階層
4	中分類	nvarchar	品詞の第 2 階層
5	小分類	nvarchar	品詞の第 3 階層
6	細分類	nvarchar	品詞の第 4 階層
7	類	nvarchar	類と品詞との対応を示す

テーブル名		ID 変換係数マスタテーブル	
説明		語彙素 ID、語形 ID、書字形 ID、発音形 ID、語彙表 ID の各 ID を別の ID に変換する際の係数マスタ	
No	フィールド名	データ型	説明
1	レベル	int	ID 階層レベル。語彙素 ID が 1(最上位)
2	ID 名	nvarchar	ID 名
3	数値 1	int	係数
4	数値 2	int	オフセット値

テーブル名		書字形構成漢字テーブル	
説明		短単位書字形テーブルの書字形に含まれる漢字を抜き出して音訓等種別、音訓を付与したテーブル(3.8・29 ページ参照)	
No	フィールド名	データ型	説明
1	書字形 ID	bigint	
2	書字形構成漢字	nvarchar	字種
3	書字形内位置	int	字種の書字形内における出現位置
4	ID	bigint	
5	書字形情報	nvarchar	
6	音訓等種別	nvarchar	
7	音訓	nvarchar	
8	精度	float	自動処理時の精度情報
9	確定	int	作業者による確認チェック
10	更新作業者	nvarchar	
11	更新日時	datetime	

【資料】

テーブル名		漢字テーブル	
説明		単漢字に関する情報を格納した表で、書字形構成漢字表と結合して利用する(3.8・29 ページ参照)	
No	フィールド名	データ型	説明
1	通し番号	int identity	
2	漢字 ID	nvarchar	
3	字種	nvarchar	
4	種類	nvarchar	
5	配当学年	int	
6	画数 1	int	
7	部首	int	
8	音訓等種別	nvarchar	
9	音訓	nvarchar	
10	音訓注記	nvarchar	
11	音訓割振	nvarchar	
12	人名制定	nvarchar	
13	日本語教育	int	
14	90 種・頻度	int	
15	90 種・音訓	nvarchar	
16	新聞・頻度	int	
17	新聞・音訓	nvarchar	
18	200 万字・頻度	int	
19	200 万字・音訓	nvarchar	
20	備考1	nvarchar	
21	備考2	nvarchar	
22	更新作業者	nvarchar	
23	更新日時	datetime	

テーブル名		出典テーブル	
説明		3.2.6(16 ページ)および資料⑥(108 ページ)参照	
No	フィールド名	データ型	説明
1	出典コード	nchar	
2	説明	varchar	
3	削除可	bit	
4	テーブル	nvarchar	

テーブル名		出現頻度テーブル	
説明		短単位テーブルにおける語彙素、語形、書字形の出現頻度	
No	フィールド名	データ型	説明
1	レベル	nvarchar	語彙素、語形、書字形のいずれか
2	ID	bigint	ID
3	内訳	nvarchar	出現頻度内訳
4	合計	int	出現頻度合計

テーブル名		短単位語形ログテーブル	
説明		3.9.4 語形削除ログ(33 ページ)参照	
No	フィールド名	データ型	説明
1	ID	int identity	
2	語形 ID	int	
3	語彙素 ID	int	
4	語形	nvarchar	
5	品詞	nvarchar	
6	活用型	nvarchar	
7	語頭変化型	nvarchar	
8	語末変化型	nvarchar	
9	出典	nvarchar	
10	DelUser	nvarchar	
11	DelDate	nchar	

テーブル名		要注意語テーブル	
説明		3.9.1 要注意語テーブル(32 ページ)参照	
No	フィールド名	データ型	説明
1	ID	int identity	
2	区分	nvarchar	
3	代表形	nvarchar	
4	代表表記	nvarchar	
5	異形態	nvarchar	
6	品詞	nvarchar	
7	活用型・その他	nvarchar	
8	接続	nvarchar	
9	注記	nvarchar	
10	削除補足	nvarchar	
11	削除	bit	

テーブル名		要注意語用例テーブル	
説明		3.9.2 要注意誤用例テーブル(32 ページ)参照	
No	フィールド名	データ型	説明
1	ID	int identity	
2	IDREF	int	
3	c	nvarchar	
4	用例	ntext	

【資料】

テーブル名		分類語彙表テーブル	
説明		3.10 分類語彙表テーブル(33 ページ)参照	
No	フィールド名	データ型	説明
1	レコード ID	int	
2	見出し番号	int	
3	レコード種別	nvarchar	
4	類	nvarchar	
5	部門	nvarchar	
6	中項目	nvarchar	
7	分類項目	nvarchar	
8	分類番号	nvarchar	
9	段落番号	nvarchar	
10	小段落番号	nvarchar	
11	語番号	nvarchar	
12	見出し	nvarchar	
13	読み	nvarchar	
14	逆読み	nvarchar	
15	見出し本体_bccwj	nvarchar	
16	読み_カタカナ	nvarchar	
17	分類語彙表番号	nvarchar	
18	読み_カタカナ_bccwj	nvarchar	
19	更新作業者	nvarchar	
20	更新日時	smalldatetime	
21	辞書データベース要登録フラグ	bit	
22	辞書データベースチェック	bit	
23	メモ	nvarchar	

テーブル名		分類語彙表関連付けテーブル	
説明		短単位語彙素テーブルと分類語彙表テーブルの中間テーブル 3.10.2 短単位語彙素テーブルとの関連付け(33 ページ)参照	
No	フィールド名	データ型	説明
1	分類語彙表番号	nvarchar	
2	語彙素 ID	int	
3	更新作業者	nvarchar	
4	更新日時	smalldatetime	

コーパスデータベース

テーブル名		短単位テーブル	
説明		4.3 短単位テーブル(38 ページ)参照	
No	フィールド名	データ型	説明
1	コーパス名	nvarchar	
2	サンプル ID	nvarchar	
3	文字開始位置	int	
4	文字終了位置	int	
5	文境界	nvarchar	
6	出現書字形	nvarchar	
7	出現発音形	nvarchar	
8	語彙素読み	nvarchar	
9	語彙素	nvarchar	
10	原文文字列	nvarchar	
11	品詞	nvarchar	
12	活用型	nvarchar	
13	活用形	nvarchar	
14	状態フラグ	nvarchar	
15	語彙表 ID	bigint	
16	語彙素細分類	nvarchar	
17	更新作業者	nvarchar	
18	更新日時	datetime	
19	連番	int	
20	メモ	ntext	
21	文字開始位置	int	
22	文字終了位置	int	
23	語種	nvarchar	
24	固定長フラグ	int	
25	可変長フラグ	int	
26	語形	nvarchar	
27	語彙素 ID	int	

テーブル名		文字テーブル	
説明		コーパスの文字開始終了位置をテーブル化したもの 4.2 コーパスデータベースのテーブル(35 ページ)参照	
No	フィールド名	データ型	説明
1	サンプル ID	nvarchar	
2	文字開始位置	int	
3	文字終了位置	int	
4	文字	nvarchar	
5	固定長フラグ	bit	BCCWJ の固定長範囲であることを示すフラグ
6	可変長フラグ	bit	BCCWJ の可変長範囲であることを示すフラグ

【資料】

テーブル名		文字修正テーブル	
説明		コーパスの文字の修正記録(BCCWJ の correction タグに相当) 4.2 コーパスデータベースのテーブル(35 ページ)参照	
No	フィールド名	データ型	説明
1	サンプル ID	nvarchar	サンプル ID
2	文字開始位置	int	文字開始位置
3	文字終了位置	int	文字終了位置
4	文字修正型	nvarchar	文字修正の種類(衍字、脱落など)
5	修正後文字	nvarchar	修正する前の文字
6	修正前文字	nvarchar	修正した後の文字
7	更新日時	smalldatetime	更新した日時
8	更新作業者	nvarchar	更新作業者名
9	メモ	ntext	更新時のメモ

テーブル名		数字テーブル	
説明		数字変換(NumTrans)箇所の記録 4.2 コーパスデータベースのテーブル(35 ページ)参照	
No	フィールド名	データ型	説明
1	サンプル ID	nvarchar	
2	文字開始位置	int	
3	文字終了位置	int	
4	出現書字形	nvarchar	
5	数字変換型	nvarchar	
6	原文文字列	nvarchar	

テーブル名		振り仮名テーブル	
説明		コーパスの文字につけられた振り仮名(BCCWJ の ruby タグに相当) 4.2 コーパスデータベースのテーブル(35 ページ)参照	
No	フィールド名	データ型	説明
1	サンプル ID	nvarchar	
2	文字開始位置	int	
3	文字終了位置	int	
4	出現書字形	nvarchar	
5	振り仮名	nvarchar	

テーブル名		タグテーブル	
説明		コーパスのタグを全て格納したもの 4.2 コーパスデータベースのテーブル(35 ページ)参照	
No	フィールド名	データ型	説明
1	サンプル ID	nvarchar	
2	出現順	int	
3	文字開始位置	int	
4	文字終了位置	int	
5	タグ	ntext	

テーブル名		語彙表テーブル	
説明		3.1(7 ページ)、3.6(25 ページ)参照	
No	フィールド名	データ型	説明
1	語彙表 ID	bigint	
2	語彙素	nvarchar	
3	語彙素読み	nvarchar	
4	類	nvarchar	
5	語彙素細分類	nvarchar	
6	語形	nvarchar	
7	品詞	nvarchar	
8	辞書登録活用型	nvarchar	
9	活用型	nvarchar	
10	活用形	nvarchar	
11	出典	nvarchar	
12	発音形(基本形)	nvarchar	
13	書字形(基本形)	nvarchar	
14	仮名形(基本形)	nvarchar	
15	出現発音形	nvarchar	
16	出現書字形	nvarchar	
17	出現仮名形	nvarchar	
18	アクセント修飾型	nvarchar	
19	状態	nvarchar	
20	語頭変化型	nvarchar	
21	語頭変化結合型	nvarchar	
22	語頭変化形	nvarchar	
23	語末変化型	nvarchar	
24	語末変化結合型	nvarchar	
25	語末変化形	nvarchar	
26	語形(基本形)	nvarchar	
27	語種	nvarchar	
28	アクセント型	nvarchar	
29	アクセント結合型	nvarchar	

テーブル名		文テーブル	
説明		全文検索用のテーブル。文境界で区切った形でコーパスデータベースの全てのテキストを格納。	
No	フィールド名	データ型	説明
1	サンプル ID	nvarchar	
2	文開始位置	int	
3	文	ntext	
4	コーパス名	nvarcha	

【資料】

テーブル名		長単位テーブル	
説明		4.4 長単位テーブルと文節テーブル(39 ページ)参照	
No	フィールド名	データ型	説明
1	サンプル ID	nvarchar	
2	長単位出現書字形	nvarchar	
3	長単位品詞	nvarchar	
4	長単位活用型	nvarchar	
5	長単位活用形	nvarchar	
6	長単位語彙素読み	nvarchar	
7	長単位語彙素	nvarchar	
8	長単位境界	nvarchar	
9	丸付き数字 1	nvarchar	
10	丸付き数字 2	nvarchar	
11	メモ	nvarchar	
12	更新作業者	nvarchar	
13	更新日時	smalldatetime	
14	長単位開始位置	int	
15	長単位終了位置	int	
16	文字開始位置	int	
17	文字終了位置	int	
18	範囲対応	int	

テーブル名		文節テーブル	
説明		4.4 長単位テーブルと文節テーブル(39 ページ)参照	
No	フィールド名	データ型	説明
1	サンプル ID	nvarchar	
2	文字開始位置	int	
3	長単位開始位置	int	
4	文節 1	nvarcha	
5	文節 2	nvarcha	
6	更新作業者	nvarcha	
7	更新日時	smalldatetime	

テーブル名		長単位語彙表テーブル	
説明		長単位用の語彙表。短単位語彙表テーブルとは異なり、辞書データベースとは連動していない。6.8.2(83 ページ)参照	
No	フィールド名	データ型	説明
1	長単位出現書字形	nvarchar	
2	長単位品詞	nvarchar	
3	長単位活用型	nvarchar	
4	長単位活用形	nvarchar	
5	長単位語彙素読み	nvarchar	
6	長単位語彙素	nvarchar	
7	ID	int	

テーブル名		検索履歴テーブル	
説明		大納言とUniDicExplorerにおける作業者の検索履歴	
No	フィールド名	データ型	説明
1	検索語	nvarchar	
2	検索方法	nvarchar	
3	検索項目	nvarchar	
4	コントロール名	nvarchar	
5	更新作業者	nvarchar	
6	更新日時	nvarchar	

テーブル名		選択肢マスタテーブル	
説明		大納言の画面内にあるコンボボックス等の選択肢マスタ	
No	フィールド名	データ型	説明
1	コントロール名	nvarchar	対応している大納言のコントロール名
2	ソート順	int	コンボボックス等における選択肢のソート順
3	文字列1	nvarchar	コンボボックス等における選択肢文字列1
4	数値1	int	コンボボックス等における選択肢数値1
5	文字列2	ntext	コンボボックス等における選択肢文字列2
6	数値2	int	コンボボックス等における選択肢数値2

サンプルデータ

① 短単位語彙素テーブル

語彙素ID	語彙素	語彙素読み	類	出典	状態	コメント	評価	原語表記	語義	語種	更新作業者	更新日時	最小単位	最小単位数
7151	亀	カメ	体	IUc						和	user1	2008/2/13 16:11	カメ/	1
7222	辛い	カライ	相	IUcr						和	user1	2008/2/13 16:11	カライ/	1
9555	着る	キル	用	IUcgpr						和	user1	2008/2/13 16:11	キル/	1
2242	一定	イッテイ	体	Icr						漢	user1	2008/2/13 16:11	イッテイ	2
5580	家	カ	接尾体	cpr						漢	user1	2008/2/13 16:11	カ	1
6801	活動	カツドウ	体	Icr						漢	user1	2008/2/13 16:11	カツドウ	2
7919	外国	ガイコク	体	Iacpr						漢	user1	2008/2/13 16:11	ガイコク	2
8329	期間	キカン	体	Iacpr						漢	user1	2008/2/13 16:11	キカン	2
9167	強化	キョウカ	体	Icr						漢	user1	2008/2/13 16:11	キョウカ	2
10988	形成	ケイセイ	体	Icr						漢	user1	2008/2/13 16:11	ケイセイ	2
11482	芸術	ゲイジュツ	体	Icr						漢	user1	2008/2/13 16:11	ゲイジュツ	2
12432	交流	コウリュウ	体	IUcr						漢	user1	2008/2/13 16:11	コウリュウ	2
12524	国際	コクサイ	体	Icr						漢	user1	2008/2/13 16:11	コクサイ	2
12836	事	コト	体	Iacgpr						和	user1	2008/2/13 16:11	コト/	1
14927	使	シ	接尾体	r						漢	user1	2008/2/13 16:11	シ	1
17256	深化	シンカ	体	IU						漢	user1	2008/2/13 16:11	シンカ	2
17803	事業	ジギョウ	体	Icr						漢	user1	2008/2/13 16:11	ジギョウ	2
18765	人	ジン	接尾体	Kacpr						漢	user1	2008/2/13 16:11	ジン	1
18917	推進	スイシン	体	Icr						漢	user1	2008/2/13 16:11	スイシン	2
19537	為る	スル	用	Lacgpr						和	user1	2008/3/11 14:13	スル/	1
20054	世界	セカイ	体	Icpr						漢	user1	2008/2/13 16:11	セカイ	2
22308	携わる	タズサワル	用	Icr						和	user1	2008/2/13 16:11	タズサワル/	1
23939	庁	チョウ	接尾体	cpr						漢	user1	2008/2/13 16:11	チョウ	1
24672	繋がる	ツナガル	用	Icr						和	user1	2008/2/13 16:11	ツナガル/	1
24874	て	テ	接助	IKacgpr						和	user1	2008/2/13 16:11	テ/	1
25355	展開	テンカイ	体	Icgr						漢	user1	2008/2/13 16:11	テンカイ	2
25826	と	ト	格助	Ikabcgpr						和	user1	2008/2/13 16:11	ト/	1
25875	等	トウ	接尾体	cr						漢	user1	2008/2/13 16:11	トウ	1
28178	に	ニ	格助	IKacgpr						和	user1	2008/2/13 16:11	ニ/	1
28455	日本	ニッポン	国	cpr						固	user1	2008/2/13 16:11	ニッポン	1
28860	ネットワーク	ネットワーク	体	Icr				network	network	外	user1	2008/2/13 16:11	ネットワーク	1
28989	の	ノ	格助	IKacgpr						和	user1	2008/2/13 16:11	ノ/	1

② 短単位語形テーブル

語形ID	語彙素ID	語形SubID	語形	品詞	活用型	語頭変化型	語頭変化結合型	語末変化型	語末変化結合型	代表性	出典	状態	コメント	評価	更新作業者	更新日時
228833	7151	1	カメ	名詞-普通名詞-一般		カ濁				1	IUc				user11	2008/3/27 17:11
228833	7151	1	カメ	名詞-普通名詞-一般		カ濁				1	IUc				user11	2008/3/27 17:11
231105	7222	1	カライ	形容詞-一般	形容詞-ライ	カ濁				1	IUcr				user11	2008/3/27 17:11
231105	7222	1	カライ	形容詞-一般	形容詞-ライ	カ濁				1	IUcr				user11	2008/3/27 17:11
231108	7222	4	カラシ	形容詞-一般	文語形容詞-ク	カ濁				0	活	k			user11	2008/3/27 17:11
231108	7222	4	カラシ	形容詞-一般	文語形容詞-ク	カ濁				0	活	k			user11	2008/3/27 17:11
305761	9555	1	キル	動詞-一般	上一段-カ行					1	IUcgpr				user11	2008/3/27 17:11
305761	9555	1	キル	動詞-一般	上一段-カ行					1	IUcgpr				user11	2008/3/27 17:11
305762	9555	2	キレル	動詞-一般	下一段-ラ行-一般					0	c				user9	2008/11/11 13:52
305762	9555	2	キレル	動詞-一般	下一段-ラ行-一般					0	c				user9	2008/11/11 13:52
305763	9555	3	キル	動詞-一般	文語上一段-カ行					0	近	k			user11	2008/9/27 14:38
305763	9555	3	キル	動詞-一般	文語上一段-カ行					0	近	k			user11	2008/9/27 14:38
71745	2242	1	イッテイ	名詞-普通名詞-サ変形状詞可能						1	Icr				user4	2008/11/6 17:49
178561	5580	1	カ	接尾辞-名詞的-一般						1	cpr				user111	2008/3/27 17:11
217633	6801	1	カツドウ	名詞-普通名詞-サ変可能						1					user111	2008/3/27 17:11
253409	7919	1	ガイコク	名詞-普通名詞-一般						1	Iacpr				user111	2008/3/27 17:11
266529	8329	1	キカン	名詞-普通名詞-一般						1	Iacpr				user111	2008/3/27 17:11
293345	9167	1	キョウカ	名詞-普通名詞-サ変可能						1	Icr				user111	2008/3/27 17:11
351617	10988	1	ケイセイ	名詞-普通名詞-サ変可能						1	Icr				user111	2008/3/27 17:11
367425	11482	1	ゲイジュツ	名詞-普通名詞-一般						1	Icr				user111	2008/3/27 17:11
397825	12432	1	コウリュウ	名詞-普通名詞-サ変可能						1	Icr				user111	2008/3/27 17:11
400769	12524	1	コクサイ	名詞-普通名詞-一般						1	Icr				user111	2008/3/27 17:11
410753	12836	1	コッ	名詞-普通名詞-一般						0	c				user111	2008/3/27 17:11
410754	12836	2	コト	名詞-普通名詞-一般		コ濁			B1WS2W	1	Iacgpr				user111	2008/3/27 17:11
477665	14927	1	シ	接尾辞-名詞的-一般						1	r				user111	2008/3/27 17:11
552193	17256	1	シンカ	名詞-普通名詞-サ変可能						1	IU				user111	2008/3/27 17:11
569697	17803	1	ジギョウ	名詞-普通名詞-一般						1	Icr				user111	2008/3/27 17:11
581026	18157	2	ジュウ	名詞-数詞			Nj	十促		1	acgpr				user111	2008/3/27 17:11
581026	18157	2	ジュウ	名詞-数詞			Nj	十促		1	acgpr				user111	2008/3/27 17:11
600481	18765	1	ジン	接尾辞-名詞的-一般						1	Kacpr				user111	2008/3/27 17:11
605345	18917	1	スイシン	名詞-普通名詞-サ変可能						1	Icr				user111	2008/3/27 17:11
625185	19537	1	スル	動詞-非自立可能	サ行変格-為ル					1	Lacgpr				user111	2008/3/27 17:11
625186	19537	2	ス	動詞-非自立可能	文語サ行変格-ス					1	CR				user111	2009/2/2 14:11
625187	19537	3	スル	動詞-非自立可能	無変化未然型-サ行変格-スル					0	b				user111	2009/2/2 14:12
641729	20054	1	セカイ	名詞-普通名詞-一般						1	Icpr				user111	2008/3/27 17:11
713857	22308	1	タズサワル	動詞-一般	五段-ラ行-一般					1	Icr				user111	2008/3/27 17:11
713858	22308	2	タズサワレル	動詞-一般	下一段-ラ行-一般					1	c				user111	2008/3/27 17:11
713859	22308	3	タズサワル	動詞-一般	文語四段-ラ行					0	活				user111	2008/3/27 17:11
766049	23939	1	チョウ	接尾辞-名詞的-一般						1	cpr				user111	2008/3/27 17:11
789505	24672	1	ツナガル	動詞-一般	五段-ラ行-一般					1	Icr				user111	2008/3/27 17:11
789506	24672	2	ツナガル	動詞-一般	文語四段-ラ行					0	活				user111	2008/3/27 17:11
795969	24874	1	テ	助詞-接続助詞						1	IKacgpr				user111	2008/3/27 17:11
795970	24874	2	デ	助詞-接続助詞						0	acgpr				user111	2008/3/27 17:11
795971	24874	3	タ	助詞-接続助詞						0	IKacgpr	c			user111	2008/3/27 17:11
795972	24874	4	ッテ	助詞-接続助詞						0	b	k			user111	2008/5/23 15:03
811361	25355	1	テンカイ	名詞-普通名詞-サ変可能						1	Icgr				user111	2008/3/27 17:11
826433	25826	1	ット	助詞-格助詞						0	bc				user111	2008/3/27 17:11
826434	25826	2	ト	助詞-格助詞						1	IKacgpr				user111	2008/3/27 17:11
826435	25826	3	トオ	助詞-格助詞						0	K				user111	2008/3/27 17:11
828001	25875	1	トウ	接尾辞-名詞的-一般						1	cr				user111	2008/3/27 17:11
901697	28178	1	ニ	助詞-格助詞						1	IKacgpr				user111	2008/6/13 17:13
901698	28178	2	ニー	助詞-格助詞						0	c	M			user111	2008/3/27 17:11
901699	28178	3	ン	助詞-格助詞						0	c	c			user111	2008/3/27 17:11
901700	28178	4	ニッ	助詞-格助詞						0	K	c			user111	2008/3/27 17:11
910561	28455	1	ニッポン	名詞-固有名詞-地名-国						1	cr				user111	2008/3/27 17:11
910562	28455	2	ニホン	名詞-固有名詞-地名-国						0	cpr				user111	2008/3/27 17:11
923521	28860	1	ネットワーク	名詞-普通名詞-一般						1	Icr				user111	2008/3/27 17:11
923522	28860	2	ネットワークス	名詞-普通名詞-一般						0	w				user111	2008/3/27 17:11

【サンプルデータ】

③ 短単位書字形テーブル

書字形ID	語形ID	書字形SubID	書字形	活用型書字形	仮名形	代表性	出典	状態	コメント	評価	更新作業者	更新日時
58581249	228833	1	かめ		カメ	0	U				user3	2008/1/7 10:00
58581249	228833	1	かめ		カメ	0	U				user2	2008/1/7 10:00
58581250	228833	2	カメ		カメ	0	U				user2	2008/1/7 10:00
58581250	228833	2	カメ		カメ	0	U				user2	2008/1/7 10:00
58581251	228833	3	亀		カメ	1	lc				user2	2008/1/7 10:00
58581251	228833	3	亀		カメ	1	lc				user2	2008/1/7 10:00
58581252	228833	4	龜		カメ	0	太	k			user2	2008/1/7 10:00
58581252	228833	4	龜		カメ	0	太	k			user2	2008/1/7 10:00
59162881	231105	1	からい	らい	カライ	0	U				user2	2008/1/7 10:00
59162881	231105	1	からい	らい	カライ	0	U				user2	2008/1/7 10:00
59162882	231105	2	辛い	一般	カライ	1	lcr				user2	2008/1/7 10:00
59162882	231105	2	辛い	一般	カライ	1	lcr				user2	2008/1/7 10:00
59162883	231105	3	辣い		カライ	0	b				user3	2008/6/12 17:19
59162883	231105	3	辣い		カライ	0	b				user3	2008/6/12 17:19
59163649	231108	1	辛し		カラシ	1	活	k			user2	2008/1/7 10:00
59163649	231108	1	辛し		カラシ	1	活	k			user2	2008/1/7 10:00
59163650	231108	2	からし		カラシ	1	活	k			user2	2008/1/7 10:00
59163650	231108	2	からし		カラシ	1	活	k			user2	2008/1/7 10:00
78274817	305761	1	きる		キル	0	U				user2	2008/1/7 10:00
78274817	305761	1	きる		キル	0	U				user2	2008/1/7 10:00
78274818	305761	2	着る		キル	1	lcgpr				user2	2008/1/7 10:00
78274818	305761	2	着る		キル	1	lcgpr				user2	2008/1/7 10:00
78275073	305762	1	着れる		キレル	0	c				user2	2008/1/7 10:00
78275073	305762	1	着れる		キレル	0	c				user2	2008/1/7 10:00
78275329	305763	1	着る		キル	0	近	k			user2	2008/1/7 10:00
78275329	305763	1	着る		キル	0	近	k			user2	2008/1/7 10:00
78275330	305763	2	衣る		キル	0	近	k			user2	2008/1/7 10:00
78275330	305763	2	衣る		キル	0	近	k			user2	2008/1/7 10:00
78275331	305763	3	きる		キル	0	近	k			kato	2008/4/15 6:49
78275331	305763	3	きる		キル	0	近	k			kato	2008/4/15 6:49
18366721	71745	1	一定		イッテイ	1	lcr				user2	2008/1/7 10:00
45711617	178561	1	家		カ	1	cpr				user2	2008/1/7 10:00
45711618	178561	2	カ		カ	0	b	Z			user3	2008/6/9 10:01
55714049	217633	1	活動		カツドウ	1	lcr				user2	2008/1/7 10:00
55714050	217633	2	カツドウ		カツドウ	0	b				user3	2008/6/19 17:11
64872705	253409	1	外国		ガイコク	1	lacpr				user2	2008/1/7 10:00
64872706	253409	2	外國		ガイコク	0	旧				user2	2008/1/7 10:00
68231425	266529	1	期間		キカン	1	lacpr				user2	2008/1/7 10:00
75096321	293345	1	強化		キョウカ	1	lcr				user2	2008/1/7 10:00
90013953	351617	1	形成		ケイセイ	1	lcr				user2	2008/1/7 10:00
94060801	367425	1	芸術		ゲイジュツ	1	lcr				user2	2008/1/7 10:00
94060802	367425	2	藝術		ゲイジュツ	0	旧				user2	2008/1/7 10:00
94060803	367425	3	ゲージュツ		ゲイジュツ	0	b				user3	2008/6/9 10:04
101843201	397825	1	交流		コウリュウ	1	lcr				user2	2008/1/7 10:00
102596865	400769	1	国際		コクサイ	1	lcr				user2	2008/1/7 10:00
102596866	400769	2	國際		コクサイ	0	旧				user2	2008/1/7 10:00
105152769	410753	1	こっ		コッ	0	c				user2	2008/1/7 10:00
105153025	410754	1	こと		コト	0	acgpr				user2	2008/1/7 10:00
105153026	410754	2	コト		コト	0	r				user2	2008/1/7 10:00
105153027	410754	3	事		コト	1	lcr				user2	2008/1/7 10:00
105153028	410754	4	丨		コト	0	Z				user2	2008/1/7 10:00
122282241	477665	1	使		シ	1	r				user2	2008/1/7 10:00
141361409	552193	1	深化		シンカ	1	lU				user2	2008/1/7 10:00
145842433	569697	1	事業		ジギョウ	1	lcr				user2	2008/1/7 10:00
148742657	581026	1	十		ジュウ	1	acgpr				user2	2008/1/7 10:00
148742657	581026	1	十		ジュウ	1	acgpr				user2	2008/1/7 10:00
148742658	581026	2	X		ジュウ	0	Z				user2	2008/1/7 10:00
148742658	581026	2	X		ジュウ	0	Z				user2	2008/1/7 10:00
148742659	581026	3	x		ジュウ	0	Z				user2	2008/1/7 10:00
148742659	581026	3	x		ジュウ	0	Z				user2	2008/1/7 10:00
148742660	581026	4	拾		ジュウ	0	t				user2	2008/1/7 10:00
148742660	581026	4	拾		ジュウ	0	t				user2	2008/1/7 10:00
148742661	581026	5	一〇		ジュッ	0	近	k			user4	2008/2/27 10:21
148742661	581026	5	一〇		ジュッ	0	近	k			user4	2008/2/27 10:21
148742662	581026	6	ジュウ		ジュウ	0	acgpr				user4	2008/12/4 10:06
148742662	581026	6	ジュウ		ジュウ	0	acgpr				user4	2008/12/4 10:06
153723137	600481	1	人		ジン	1	acpr				user2	2008/1/7 10:00
153723138	600481	2	ジン		ジン	0	K				user2	2008/1/7 10:00
154968321	605345	1	推進		スイシン	1	lcr				user2	2008/1/7 10:00
160047361	625185	1	する		スル	1					user2	2008/1/7 10:00
160047362	625185	2	為る		スル	0					user2	2008/1/7 10:00
160047363	625185	3	仕る		スル	0					user2	2008/1/7 10:00
160047364	625185	4	爲る		スル	0	近	k			user6	2008/4/25 12:11
160047618	625186	2	爲		ス	0	近	k		○	user2	2008/1/7 10:00
160047619	625186	3	為		ス	0	近	k			user2	2008/1/7 10:00
160047620	625186	4	す		ス	0	近				user1	2008/11/11 10:43
160047873	625187	1	せえ		セエ	1	b		関西方言		user1	2008/4/3 16:25
164282625	641729	1	世界		セカイ	1	lcpr				user2	2008/1/7 10:00
164282626	641729	2	せかい		セカイ	0	b			児童	user1	2008/6/17 14:49
164282627	641729	3	セカイ		セカイ	0	Z				user1	2008/6/17 14:49
182747393	713857	1	たずさわる		タズサワル	0	r				user2	2008/1/7 10:00
182747394	713857	2	携る		タズサワル	0					user2	2008/1/7 10:00
182747395	713857	3	携わる		タズサワル	1	lcr				user2	2008/1/7 10:00
182747396	713857	4	たづさわる		タズサワル	0	w				user2	2008/1/7 10:00
182747649	713858	1	携われる		タズサワレル	1	c				user2	2008/1/7 10:00
182747650	713858	2	たずさわれる		タズサワレル	0	Z				ogiso	2009/2/8 1:28
182747905	713859	1	たずさわる		タズサワル	1	活	k			user2	2008/1/7 10:00
182747907	713859	3	携る		タズサワル	1	活	k			user2	2008/1/7 10:00
182747908	713859	4	携わる		タズサワル	1	活	k			user2	2008/1/7 10:00
182747909	713859	5	携はる		タズサワル	0	近	k			sunaga	2008/11/7 9:52
182747910	713859	6	たづさはる		タヅサハル	0	太	k			ogiso	2009/2/7 20:12
196108545	766049	1	庁		チョウ	1	cpr				user2	2008/1/7 10:00
196108546	766049	2	廳		チョウ	0	近				user2	2008/1/7 10:00
202113281	789505	1	つながる		ツナガル	0	r				user2	2008/1/7 10:00
202113282	789505	2	繋がる		ツナガル	1	lc				user2	2008/1/7 10:00
202113283	789505	3	繋がる		ツナガル	0	w				user2	2008/1/7 10:00
202113284	789505	4	繋る		ツナガル	1	b				user3	2008/6/18 13:56
202113537	789506	1	つながる		ツナガル	1	活	k			user2	2008/1/7 10:00
202113538	789506	2	繋がる		ツナガル	1	活	k			user2	2008/1/7 10:00
202113539	789506	3	繋がる		ツナガル	1	活	k			user2	2008/1/7 10:00
202113540	789506	4	繋る		ツナガル	0	近	k			user9	2008/4/1 12:05
233103617	910561	1	ニッポン		ニッポン	0	r				user2	2008/1/7 10:00
233103618	910561	2	日本		ニッポン	1	cr				user2	2008/1/7 10:00
233103619	910561	3	にっぽん		ニッポン	0	b				user1	2008/4/3 10:53
233103620	910561	4	日(本		ニッポン	0	y	Z			user3	2008/10/20 14:23
233103873	910562	1	ニホン		ニホン	0	r				user2	2008/1/7 10:00
233103874	910562	2	日本		ニホン	0	cpr				user2	2008/1/7 10:00
236421377	923521	1	ネットワーク		ネットワーク	1	lcr				user2	2008/1/7 10:00
236421378	923521	2	NETWORK		ネットワーク	0	w				user2	2008/1/7 10:00
236421379	923521	3	Network		ネットワーク	0	b	仮			user11	2009/2/12 14:16
236421633	923522	1	ネットワークス		ネットワークス	1	w				user2	2008/1/7 10:00

④ 短単位発音形テーブル

発音形ID	語形ID	発音形SubID	発音形	活用型発音形	アクセント型	アクセント結合型	代表性	出典	アクセント型出典	状態	コメント	評価	更新作業者	更新日時
58581249	228833	1	カメ		1	C3	1	IUc					user1	2008/2/13 16:14
58581249	228833	1	カメ		1	C3	1	IUc					user1	2008/2/13 16:14
59162881	231105	1	カライ		2	C1	1	IUcr					user1	2008/2/13 16:14
59162881	231105	1	カライ		2	C1	1	IUcr					user1	2008/2/13 16:14
59163649	231108	1	カラシ		1	C1	1	活					user1	2008/2/13 16:14
59163649	231108	1	カラシ		1	C1	1	活					user1	2008/2/13 16:14
78274817	305761	1	キル		0	C4	1	IUcgpr					user1	2008/2/13 16:14
78274817	305761	1	キル		0	C4	1	IUcgpr					user1	2008/2/13 16:14
78275073	305762	1	キレル		0	C2	0	c					user1	2008/2/13 16:14
78275073	305762	1	キレル		0	C2	0	c					user1	2008/2/13 16:14
78275329	305763	1	キル		0	C4	0	近		k			user1	2008/1/7 10:01
78275329	305763	1	キル		0	C4	0	近		k			user1	2008/1/7 10:01
18366721	71745	1	イッテー		0	C2	1	Icr					user1	2008/2/13 16:14
45711617	178561	1	カ			C4	1	cpr					user1	2008/2/13 16:14
55714049	217633	1	カンドー		0	C2	1	Icr					user1	2008/2/13 16:14
64872705	253409	1	ガイコク		0	C2	1	Iacpr					user1	2008/2/13 16:14
68231425	266529	1	キカン		1,2	C1	1	Iacpr					user1	2008/2/13 16:14
75096321	293345	1	キョーカ		1,0	C1	1	Icr	D=1N-1				user1	2008/2/13 16:14
90013953	351617	1	ケーセー		0	C2	1	Icr					user1	2008/2/13 16:14
94060801	367425	1	ゲージュツ		0,1	C2	1	Icr	D=1N-1				user1	2008/2/13 16:14
101843201	397825	1	コーリュー		0	C2	1	Icr					user1	2008/2/13 16:14
102596865	400769	1	コクサイ		0	C2	1	Icr					user1	2008/2/13 16:14
105152769	410753	1	コッ		2	C3	0	c					user1	2008/2/13 16:14
105153025	410754	1	コト		2	C3	1	Iacgpr					user1	2008/2/13 16:14
122282241	477665	1	シ			C3	1	r					user1	2008/2/13 16:14
141361409	552193	1	シンカ		1	C1	1	IU					user1	2008/2/13 16:14
145842433	569697	1	ジギョー		1	C1	1	Icr					user1	2008/2/13 16:14
148742657	581026	1	ジュー		1	C3	1	acgpr					user1	2008/2/13 16:14
148742657	581026	1	ジュー		1	C3	1	acgpr					user1	2008/2/13 16:14
153723137	600481	1	ジシ			C3	1	Kacpr					user1	2008/2/13 16:14
154968321	605345	1	スイシン		0	C2	1	Icr					user1	2008/2/13 16:14
160047361	625185	1	スル		0	C5	1	Lacgpr					user1	2008/2/13 16:14
160047617	625186	1	ス		1	C4	0						user1	2008/2/13 16:14
160047873	625187	1	セー		1	C3	1	b					user1	2008/4/3 16:26
164282625	641729	1	セカイ		1,2	C1	1	Icpr	D=1N-1				user1	2008/2/13 16:14
182747393	713857	1	タズサワル		4	C1	1	Icr					user1	2008/2/13 16:14
182747649	713858	1	タズサワレル		5	C1	1	c					user1	2008/1/7 10:01
182747905	713859	1	タズサワル		4	C1	1	活					user1	2008/2/13 16:14
196108545	766049	1	チョー			C3	1	cpr					user1	2008/2/13 16:14
202113281	789505	1	ツナガル		0	C2	1	Icr					user1	2008/2/13 16:14
202113537	789506	1	ツナガル		0	C2	1	活					user1	2008/2/13 16:14
203768065	795969	1	テ			動詞%F1,形容詞%F2@-1	1	IKacgpr					user1	2008/2/13 16:14
203768321	795970	1	デ			動詞%F1,形容詞%F2@-1	0	acgpr					user1	2008/2/13 16:14
203768577	795971	1	タ				1	IKacgpr		c			user1	2008/3/6 13:52
203768833	795972	1	ッテ				1	b					user1	2008/5/23 15:03
207708417	811361	1	テンカイ		0	C2	1	Icgr					user1	2008/2/13 16:14
211566849	826433	1	ット			名詞%F1,動詞%F1,形容詞%F2@-1	0	bc					user1	2008/2/13 16:14
211567105	826434	1	ト			名詞%F1,動詞%F1,形容詞%F2@-1	1	IKacgpr					user1	2008/2/13 16:14
211567361	826435	1	トー			名詞%F1,動詞%F1,形容詞%F2@-1	0	K					user1	2008/2/13 16:14
211968257	828001	1	トー			C1	1	cr					user1	2008/2/13 16:14
230834433	901697	1	ニ			名詞%F1	1	IKacgpr					user1	2008/2/13 16:14
230834689	901698	1	ニー			名詞%F1	0	c		M			user1	2008/1/7 10:01
230834945	901699	1	ン			名詞%F1	0	c					user1	2008/2/13 16:14
230835201	901700	1	ニッ			名詞%F1	0	K					user1	2008/2/13 16:14
233103617	910561	1	ニッポン		3		1	cr					user1	2008/2/13 16:14
233103873	910562	1	ニホン		2		0	cpr					user1	2008/2/13 16:14
236421377	923521	1	ネットワーク		4	C1	1	Icr					user1	2008/2/13 16:14
236421633	923522	1	ネットワークス		4	C1	1	w					user1	2008/2/13 16:14

【サンプルデータ】

⑤ 書字形構成漢字テーブル

書字形ID	書字形構成漢字	書字形内位置	ID	書字形情報	音訓等種別	音訓	精度	確定	更新作業者	更新日時
18366721	一	1	587735073	一定:イッテイ	音	イツ	0.5	1	user5	2007/12/21 15:03
18366721	定	2	587735074	一定:イッテイ	音	テイ	1	1	user5	2007/12/21 15:03
45711617	家	1	1462771745	家:カ	音	カ	1	1	user5	2007/9/3 16:09
55714049	活	1	1782849569	活動:カツドウ	音	カツ	1	1	user5	2007/9/6 9:43
55714049	動	2	1782849570	活動:カツドウ	音	ドウ	1	1	user5	2007/9/6 9:43
64872705	外	1	2075926561	外国:ガイコク	音	ガイ	1	1	user5	2007/9/6 11:04
64872705	国	2	2075926562	外国:ガイコク	音	コク	1	1	user5	2007/9/6 11:04
68231425	期	1	2183405601	期間:キカン	音	キ	1	1	user5	2007/9/6 11:28
68231425	間	2	2183405602	期間:キカン	音	カン	1	1	user5	2007/9/6 11:28
75096321	強	1	2403082273	強化:キョウカ	音	キョウ	1	1	user5	2007/9/6 13:22
75096321	化	2	2403082274	強化:キョウカ	音	カ	1	1	user5	2007/9/6 13:22
90013953	形	1	2880446497	形成:ケイセイ	音	ケイ	1	1	user5	2007/9/6 16:24
90013953	成	2	2880446498	形成:ケイセイ	音	セイ	1	1	user5	2007/9/6 16:24
94060801	芸	1	3009945633	芸術:ゲイジュツ	音	ゲイ	1	1	user5	2007/9/6 17:14
94060801	術	2	3009945634	芸術:ゲイジュツ	音	ジュツ	1	1	user5	2007/9/6 17:14
101843201	交	1	3258982433	交流:コウリュウ	音	コウ	1	1	user5	2007/9/10 10:35
101843201	流	2	3258982434	交流:コウリュウ	音	リュウ	1	1	user5	2007/9/10 10:35
102596865	国	1	3283099681	国際:コクサイ	音	コク	1	1	user5	2007/9/10 10:43
102596865	際	2	3283099682	国際:コクサイ	音	サイ	1	1	user5	2007/9/10 10:43
122282241	使	1	3913031713	使:シ	音	シ	1	1	user5	2007/9/10 14:58
141361409	深	1	4523565089	深化:シンカ	音	シン	1	1	user5	2007/9/11 12:12
141361409	化	2	4523565090	深化:シンカ	音	カ	1	1	user5	2007/9/11 12:12
145842433	事	1	4666957857	事業:ジギョウ	音	ジ	0.9	1	user5	2007/10/20 18:10
145842433	業	2	4666957858	事業:ジギョウ	音	ギョウ	1	1	user5	2007/10/20 18:10
153723137	人	1	4919140385	人:ジン	音	ジン	1	1	user5	2007/9/11 14:44
154968321	推	1	4958986273	推進:スイシン	音	スイ	1	1	user5	2007/9/11 14:58
154968321	進	2	4958986274	推進:スイシン	音	シン	1	1	user5	2007/9/11 14:58
164282625	世	1	5257044001	世界:セカイ	音	セ	1	1	user5	2007/9/13 9:21
164282625	界	2	5257044002	世界:セカイ	音	カイ	1	1	user5	2007/9/13 9:21
182747395	携	1	5847916641	携わる:タズサワル	訓	たずさわる	1	1	user5	2008/1/30 14:32
196108545	庁	1	6275473441	庁:チョウ	音	チョウ	1	1	user5	2007/9/13 17:12
207708417	展	1	6646669345	展開:テンカイ	音	テン	1	1	user5	2007/9/14 11:14
207708417	開	2	6646669346	展開:テンカイ	音	カイ	1	1	user5	2007/9/14 11:14
211968257	等	1	6782984225	等:トウ	音	トウ	1	1	user5	2007/9/14 11:53
233103618	日	1	7459315777	日本:ニッポン	国	ニッポン	1	1	user5	2007/9/18 9:50
233103618	本	2	7459315778	日本:ニッポン	国	ニッポン	1	1	user5	2007/9/18 9:50
244646145	発	1	7828676641	発信:ハッシン	音	ハツ	0.5	1	user5	2007/7/5 10:39
244646145	信	2	7828676642	発信:ハッシン	音	シン	1	1	user5	2007/7/5 10:39
258540033	人	1	8273281057	人々:ヒトビト	訓	ひと	1	1	user5	2007/9/18 14:43
274841857	文	1	8794939425	文化:ブンカ	音	ブン	1	1	user5	2007/9/19 9:58
274841857	化	2	8794939426	文化:ブンカ	音	カ	1	1	user5	2007/9/19 9:58
308642049	目	1	9876545569	目的:モクテキ	音	モク	1	1	user5	2007/9/19 17:03
308642049	的	2	9876545570	目的:モクテキ	音	テキ	1	1	user5	2007/9/19 17:03
326787329	理	1	10457194529	理解:リカイ	音	リ	1	1	user5	2007/9/20 11:38
326787329	解	2	10457194530	理解:リカイ	音	カイ	1	1	user5	2007/9/20 11:38

⑥ 漢字テーブル

通し番号	漢字ID	字種	種類	配当学年	画数1	画数2	部首	音訓等種別	音訓	音訓注記	音訓制限	人名制定	日本語教育	90種・頻度	90種・音訓	新聞・頻度	新聞・音訓	200万字・	200万字・	備考1	備考2	更新作業者	更新日時
23941	113160_1-36-74	定	教育	3				特殊訓	ぶじょう		表外											user8	2008/12/9 12:02
3135	113160_1-36-74	定	教育	3				訓	さだか		高		20	456		2763	39	1884	8				
3136	113160_1-36-74	定	教育	3				訓	さだまる		小		20	456	1	2763		1884					
3137	113160_1-36-74	定	教育	3				訓	さだめる		小		20	456	19	2763		1884	15				
3133	113160_1-36-74	定	教育	3	8	40	宀	音	ジョウ		小		20	456	23	2763	44	1884					
3134	113160_1-36-74	定	教育	3				音	テイ		小		20	456	400	2763	2658	1884	1753				
317	101240_1-18-40	家	教育	2				付表	おもや		高			876		2293		1450					
6890	101240_1-18-40	家	教育	2				特殊訓	あひる														
16670	101240_1-18-40	家	教育	2				特殊訓	えふね		表外											user8	2007/12/20 16:20
314	101240_1-18-40	家	教育	2				訓	いえ		小		30	876	209	2293	325	1450	244				
315	101240_1-18-40	家	教育	2				訓	うち		表外			876		2293		1450	10				
316	101240_1-18-40	家	教育	2				訓	や		小		30	876	22	2293	86	1450	41				
312	101240_1-18-40	家	教育	2	10	40	宀	音	カ		小		30	876	515	2293	1727	1450	881				
313	101240_1-18-40	家	教育	2				音	ケ		小		30	876	93	2293	81	1450	78				
23649	102010_1-19-72	活	教育	2				特殊訓	いき		表外											user8	2008/10/17 9:50
6893	102010_1-19-72	活	教育	2				特殊訓	うど														
23691	102010_1-19-72	活	教育	2				特殊訓	しむちょん		表外											user8	2008/10/17 14:52
23315	102010_1-19-72	活	教育	2				特殊訓	たつき		表外											user8	2008/9/19 9:50
22489	102010_1-19-72	活	教育	2				特殊訓	なりわい		表外											user8	2008/6/6 14:35
507	102010_1-19-72	活	教育	2				訓	いかす		表外			419		1209		746	12				
508	102010_1-19-72	活	教育	2				訓	いきる		表外			419		1209		746	2				
509	102010_1-19-72	活	教育	2				訓	いける		表外			419		1209		746	1				
506	102010_1-19-72	活	教育	2	9	85	舌	音	カツ		小		20	419	405	1209	1205	746	699				
16900	114080_1-38-16	動	教育	3				特殊訓	とよむ		表外											user8	2008/1/18 9:49
9896	114080_1-38-16	動	教育	3				特殊訓	どよむ		表外											user8	2007/4/27 14:58
9898	114080_1-38-16	動	教育	3				特殊訓	どよめく		表外											user8	2007/4/27 14:58
12777	114080_1-38-16	動	教育	3				特殊訓	みじろぎ		表外											user8	2007/6/4 9:55
16589	114080_1-38-16	動	教育	3				特殊訓	みじろぐ		表外											user8	2007/12/14 16:31
3330	114080_1-38-16	動	教育	3				訓	うごかす		小		30	521	27	2287	28	1450	47				
3331	114080_1-38-16	動	教育	3				訓	うごく		小		30	521	96	2287	396	1450	180				
15845	114080_1-38-16	動	教育	3				訓	やや		表外											user8	2007/12/7 9:51
3329	114080_1-38-16	動	教育	3	11	19	力	音	ドウ		小		30	521	396	2287	1862	1450	1096				
12170	101690_1-19-16	外	教育	2				特殊訓	ういろう		表外											user8	2007/5/21 17:10
22475	101690_1-19-16	外	教育	2				特殊訓	けれん		表外											user8	2008/6/6 14:21
16943	101690_1-19-16	外	教育	2				特殊訓	それる		表外											user8	2008/1/18 14:53
13289	101690_1-19-16	外	教育	2				特殊訓	とつくに		表外											user8	2007/6/11 13:17
23311	101690_1-19-16	外	教育	2				特殊訓	どうけ		表外											user8	2008/9/19 9:47
22827	101690_1-19-16	外	教育	2				特殊訓	よそひと		表外											user8	2008/7/11 12:09
435	101690_1-19-16	外	教育	2				訓	そと		小		40	556	82	1850	112	1036	91				
11554	101690_1-19-16	外	教育	2				訓	そらす		表外											user8	2007/5/17 16:23
14851	101690_1-19-16	外	教育	2				訓	と		表外											user2	2007/9/21 13:18
436	101690_1-19-16	外	教育	2				訓	はずす		小		40	556	9	1850	1	1036	38				
437	101690_1-19-16	外	教育	2				訓	はずれる		小		40	556	4	1850	4	1036	13				
438	101690_1-19-16	外	教育	2				訓	ほか		小		40	556	27	1850	2	1036	2				
433	101690_1-19-16	外	教育	2	5	36	夕	音	ガイ		小		40	556	418	1850	1679	1036	711				
434	101690_1-19-16	外	教育	2				音	ゲ		中		40	556	3	1850	21	1036	7				
13290	105940_1-25-81	国	教育	2				特殊訓	とつくに		表外											user8	2007/6/11 13:17
1413	105940_1-25-81	国	教育	2				訓	くに		小		40	1057	128	7723	661	1774	138				
1412	105940_1-25-81	国	教育	2	8	31	口	音	コク		小		40	1057	757	7723	5334	1774	1144				
22387	102880_1-20-92	期	教育	3				特殊訓	とき		表外											user8	2008/6/6 11:50
669	102880_1-20-92	期	教育	3	12	74	月	音	キ		小		20	475	469	1825	1822	943	875				
670	102880_1-20-92	期	教育	3				音	ゴ	◇	高		20	475	5	1825	3	943					

⑦ 語彙表テーブル

語彙素ID	語彙素	語彙素読み	頻	語彙素細分類	品詞	辞書意味活用型	活用型	活用形	出典	発音形（基本形）	書字形（基本形）	仮名形（基本形）	出現形発音形	出現形書字形	出現形仮名形	アクセント修飾型	状態	語種変化形	語種変化結合	語頭変化形	語末変化型	語末変化結合	語末変化形	語種	語種形（基本形）	アクセント型	アクセント結合型
195660536185344	魚	カメ	体	カメ	名詞-普通名詞-一般			語幹-一般	U	カメ	かめ	カメ	カメ	かめ	カメ		カ濁			基本形				カメ	和	3	C3
195660536193536	亀	カメ	体	カメ	名詞-普通名詞-一般			語幹-一般	U	カメ	かめ	カメ	カメ	かめ	カメ		カ濁			基本形				カメ	和	3	C3
195660536973776	亀	カメ	体	カメ	名詞-普通名詞-一般			語幹-一般	U	カメ	かめ	カメ	カメ	かめ	カメ		カ濁			基本形				カメ	和	3	C3
195660536974976	亀	カメ	体	カメ	名詞-普通名詞-一般			語幹-一般	U	カメ	かめ	カメ	カメ	かめ	カメ		カ濁			基本形				カメ	和	3	C3
195660603294208	亀	カメ	体	カメ	名詞-普通名詞-一般			語幹-一般	U	カメ	かめ	カメ	カメ	かめ	カメ		カ濁			基本形				カメ	和	3	C3
195660603302400	亀	カメ	体	カメ	名詞-普通名詞-一般			語幹-一般	U	カメ	かめ	カメ	カメ	かめ	カメ		カ濁			基本形				カメ	和	3	C3
195660603848640	亀	カメ	体	カメ	名詞-普通名詞-一般			語幹-一般	U	カメ	かめ	カメ	カメ	かめ	カメ		カ濁			基本形				カメ	和	3	C3
195660636856832	亀	カメ	体	カメ	名詞-普通名詞-一般			語幹-一般	U	カメ	かめ	カメ	カメ	かめ	カメ		カ濁			基本形				カメ	和	3	C3
195817686578401	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	語幹-一般		U	カライ	から	カラ	カライ	から	カラ		カ濁			基本形				カライ	和	2	C1
195817686578465	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686578497	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-一般		U	カライ	からく	カラク	カライ	からく	カラク	M1@0	カ濁			基本形				カライ	和	2	C1
195817686578499	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-ウ音便		U	カライ	からう	カラウ	カライ	からう	カロウ		カ濁			基本形				カライ	和	2	C1
195817686578500	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-促音便		U	カライ	からっ	カラッ	カライ	からっ	カラッ	M2@2	カ濁			基本形				カライ	和	2	C1
195817686578531	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	終止形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686578539	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	終止形-一般		U	カライ	からい	カライ	カライ	からい	カライ		k	カ濁		基本形				カライ	和	2	C1
195817686578543	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686578561	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-一般		U	カライ	からけれ	カラケレ	カライ	かれぇ	カラケレ	M2@2	カ濁			基本形				カライ	和	2	C1
195817686578593	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-融合		U	カライ	からけりゃ	カラケリャ	カライ	かれぇ	カラケリャ	M2@2	カ濁			基本形				カライ	和	2	C1
195817686578595	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-融合		U	カライ	からきゃ	カラキャ	カライ	かれぇ	カラキャ	M2@1	カ濁			基本形				カライ	和	2	C1
195817686578664	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-促音便		U	カライ	からっ	カラッ	カライ	からっ	カラッ	M2@2	カ濁			基本形				カライ	和	2	C1
195817686578693	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	語幹-一般		U	カライ	から	カラ	カライ	から	カラ		カ濁			基本形				カライ	和	2	C1
195817686586657	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686586659	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-一般		U	カライ	からく	カラク	カライ	からく	カラク	M1@0	カ濁			基本形				カライ	和	2	C1
195817686586683	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-ウ音便		U	カライ	からう	カラウ	カライ	からう	カロウ		カ濁			基本形				カライ	和	2	C1
195817686586692	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-促音便		U	カライ	からっ	カラッ	カライ	からっ	カラッ	M2@2	カ濁			基本形				カライ	和	2	C1
195817686586723	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	終止形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686586731	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	終止形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686586734	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686586735	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686586753	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817686586785	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-一般		U	カライ	からけれ	カラケレ	カライ	かれぇ	カラケレ	M2@2	カ濁			基本形				カライ	和	2	C1
195817686586786	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-融合		U	カライ	からけりゃ	カラケリャ	カライ	かれぇ	カラケリャ	M2@2	カ濁			基本形				カライ	和	2	C1
195817686586787	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-融合		U	カライ	からきゃ	カラキャ	カライ	かれぇ	カラキャ	M2@1	カ濁			基本形				カライ	和	2	C1
195817686586856	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-促音便		U	カライ	からっ	カラッ	カライ	からっ	カラッ	M2@2	カ濁			基本形				カライ	和	2	C1
195817690132832	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	語幹-一般		U	カライ	から	カラ	カライ	から	カラ		カ濁			基本形				カライ	和	2	C1
195817690132963	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690132971	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	終止形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690132974	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	終止形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690132975	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690132983	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690133025	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-一般		U	カライ	からけれ	カラケレ	カライ	かれぇ	カラケレ	M2@2	カ濁			基本形				カライ	和	2	C1
195817690133026	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-融合		U	カライ	からけりゃ	カラケリャ	カライ	かれぇ	カラケリャ	M2@2	カ濁			基本形				カライ	和	2	C1
195817690133027	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-融合		U	カライ	からきゃ	カラキャ	カライ	かれぇ	カラキャ	M2@1	カ濁			基本形				カライ	和	2	C1
195817690133096	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-促音便		U	カライ	からっ	カラッ	カライ	からっ	カラッ	M2@2	カ濁			基本形				カライ	和	2	C1
195817690141025	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	語幹-一般		U	カライ	から	カラ	カライ	から	カラ		カ濁			基本形				カライ	和	2	C1
195817690141089	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690141091	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-一般		U	カライ	からく	カラク	カライ	からく	カラク	M1@0	カ濁			基本形				カライ	和	2	C1
195817690141121	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-ウ音便		U	カライ	からう	カラウ	カライ	からう	カロウ		カ濁			基本形				カライ	和	2	C1
195817690141123	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-促音便		U	カライ	からっ	カラッ	カライ	からっ	カラッ	M2@2	カ濁			基本形				カライ	和	2	C1
195817690141124	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	終止形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690141155	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	終止形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690141166	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690141167	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連体形-一般		U	カライ	からい	カライ	カライ	からい	カライ		カ濁			基本形				カライ	和	2	C1
195817690141185	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-一般		U	カライ	からけれ	カラケレ	カライ	かれぇ	カラケレ	M2@2	カ濁			基本形				カライ	和	2	C1
195817690141217	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-融合		U	カライ	からけりゃ	カラケリャ	カライ	かれぇ	カラケリャ	M2@2	カ濁			基本形				カライ	和	2	C1
195817690141218	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	仮定形-融合		U	カライ	からきゃ	カラキャ	カライ	かれぇ	カラキャ	M2@1	カ濁			基本形				カライ	和	2	C1
195817690141219	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-促音便		U	カライ	からっ	カラッ	カライ	からっ	カラッ	M2@2	カ濁			基本形				カライ	和	2	C1
195817690141288	羊い	カライ	相	形容詞-一般	形容詞-ライ	形容詞	連用形-促音便		U	カライ	からっ	カラッ	カライ	からっ	カラッ	M2@2	カ濁			基本形				カライ	和	2	C1

【サンプルデータ】

⑧ 短単位テーブル

コーパス名	サンプルID	文字開始位置	文字終了位置	文境界	出現形	出現発音形	語彙素読み	語彙素	原文文字列	品詞	解析活用型	活用形	状態フラグ	語彙素ID	語彙素細分類	更新作業者	更新日	連濁	メモ	文節開始位置	文節終了位置	語頭	固定長フラグ	可変長フラグ	語形	語彙素ID
OW core	OW6X 00000	10	20	B	1	イチ	一	一	一	名詞-数詞				5635084335527696			10			10	20	0	1		イチ	2050
OW core	OW6X 00000	20	30	I						空白							20			20	30	0	1			23
OW core	OW6X 00000	30	50	I	日本	ニッポン	日本	日本	日本	名詞-固有名詞-地名-国				6330815488512			30			20	50	0	1		ニッポン	28455
OW core	OW6X 00000	50	70	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				7821659499274752			50			50	70	0	1		ブンカ	28455
OW core	OW6X 00000	70	80	I	の	ノ	の	の	の	助詞-格助詞				9222162401600000			70			70	80	0	1		ノ	33550
OW core	OW6X 00000	80	100	I	発展	ハッテン	発展	発展	発展	名詞-普通名詞-サ変可能				8209898243660844 16			80			80	100	0	1		ハッテン	28864
OW core	OW6X 00000	100	110	I	に	ニ	に	に	に	助詞-格助詞				7745518285496832			100			100	110	0	1		ニ	28178
OW core	OW6X 00000	110	130	I	よる	ヨル	よる	寄る	寄る	動詞-一般	五段-ラ行-一般	連体形-一般		1086345838396281 7			110			110	130	0	1		ヨル	38921
OW core	OW6X 00000	130	150	I	国際	コクサイ	国際	国際	国際	名詞-普通名詞-一般				3442579530195456			130			130	150	0	1		コクサイ	12524
OW core	OW6X 00000	150	170	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				9222162401600000			150			150	170	0	1		ブンカ	33550
OW core	OW6X 00000	170	190	I	交流	コーリュー	交流	交流	交流	名詞-普通名詞-サ変可能				3417290782756608			170			170	190	0	1		コウリュウ	12432
OW core	OW6X 00000	190	200	I	の	ノ	の	の	の	助詞-格助詞				7980444286028416			190			190	200	0	1		ノ	28989
OW core	OW6X 00000	200	220	I	推進	スイシン	推進	推進	推進	名詞-普通名詞-サ変可能				5139873898238448			200			200	220	0	1		スイシン	18917
OW core	OW6X 00000	220	240	B	、	、	、	、	、	補助記号-読点				1265300734884224			130			220	240	0	1		、	46
OW core	OW6X 00000	240	250	I	イチ	イチ	一	一	一	名詞-数詞				5635084335527296			150			230	250	0	1		イチ	2050
OW core	OW6X 00000	250	270	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				9222162401600000			170			250	270	0	1		ブンカ	33550
OW core	OW6X 00000	270	280	I	チョー	チョウ	庁	庁	庁	接尾辞-名詞的-一般				6580310837861216			180			270	280	0	1		チョウ	23939
OW core	OW6X 00000	280	300	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				9222162401600000			190			280	300	0	1		ブンカ	33550
OW core	OW6X 00000	300	320	I	交流	コーリュー	交流	交流	交流	名詞-普通名詞-サ変可能				3417290782756608			200			300	320	0	1		コウリュウ	12432
OW core	OW6X 00000	320	330	I	会	カイ	会	会	会	接尾辞-名詞的-一般				4103111140581888			210			320	330	0	1		カイ	14927
OW core	OW6X 00000	330	350	I	推進	スイシン	推進	推進	推進	名詞-普通名詞-サ変可能				4893860000952832			220			330	350	0	1		スイシン	18917
OW core	OW6X 00000	350	360	B	、	、	、	、	、	補助記号-読点				5635084335527296			230			350	360	0	1		、	46
OW core	OW6X 00000	360	370	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				6330815488512			240			360	370	0	1		ブンカ	33550
OW core	OW6X 00000	370	390	I	チョー	チョウ	庁	庁	庁	接尾辞-名詞的-一般				6580310837861216			250			370	390	0	1		チョウ	23939
OW core	OW6X 00000	390	400	I	ブンカ	ブンカ	文化	文化	文化	名詞-普通名詞-一般				9222162401600000			260			390	400	0	1		ブンカ	33550
OW core	OW6X 00000	400	420	I	コーリュー	コウリュウ	交流	交流	交流	名詞-普通名詞-サ変可能				3417290782756608			270			400	420	0	1		コウリュウ	12432
OW core	OW6X 00000	420	440	I	会	カイ	会	会	会	接尾辞-名詞的-一般				4103111140581888			280			420	440	0	1		カイ	14927
OW core	OW6X 00000	440	450	I	推進	スイシン	推進	推進	推進	名詞-普通名詞-サ変可能				4893860000952832			300			440	450	0	1		スイシン	18917
OW core	OW6X 00000	450	470	I	委員	イイン	委員	委員	委員	名詞-普通名詞-一般				4893860000952832			310			450	470	0	1		イイン	14927
OW core	OW6X 00000	470	480	B	、	、	、	、	、	補助記号-読点				6330815488512			320			470	480	0	1		、	23
OW core	OW6X 00000	480	500	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				9222162401600000			320			480	500	0	1		ブンカ	33550
OW core	OW6X 00000	500	510	I	庁	チョー	庁	庁	庁	接尾辞-名詞的-一般				6580310837861216			340			500	510	0	1		チョウ	23939
OW core	OW6X 00000	510	530	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				9222162401600000			340			510	530	0	1		ブンカ	33550
OW core	OW6X 00000	530	550	I	交流	コーリュー	交流	交流	交流	名詞-普通名詞-サ変可能				3417290782756608			350			530	550	0	1		コウリュウ	12432
OW core	OW6X 00000	550	560	I	会	カイ	会	会	会	接尾辞-名詞的-一般				4103111140581888			360			550	560	0	1		カイ	14927
OW core	OW6X 00000	560	580	I	推進	スイシン	推進	推進	推進	名詞-普通名詞-サ変可能				4893860000952832			370			560	580	0	1		スイシン	17803
OW core	OW6X 00000	580	590	I	は	ワ	は	は	は	助詞-係助詞				8059703733133824			380			580	590	0	1		ハ	29321
OW core	OW6X 00000	590	600	I	、	、	、	、	、	補助記号-読点				1375251897600 0			390			590	600	0	1		、	50
OW core	OW6X 00000	600	620	I	芸術	ゲージュツ	芸術	芸術	芸術	名詞-普通名詞-一般				3156156751159808			400			600	620	0	1		ゲイジュツ	11482
OW core	OW6X 00000	620	630	I	家	カ	家	家	家	接尾辞-名詞的-一般				1533827344376320			410			620	630	0	1		カ	5580
OW core	OW6X 00000	630	640	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				1375251897600 0			420			630	640	0	1		ブンカ	50
OW core	OW6X 00000	640	660	I	人	ジン	人	人	人	接尾辞-名詞的-一般				9222162401600000			430			640	660	0	1		ジン	33550
OW core	OW6X 00000	660	670	I	等	トー	等	等	等	接尾辞-名詞的-一般				5158092547432960			440			660	670	0	1		トウ	18765
OW core	OW6X 00000	670	680	I	、	、	、	、	、	補助記号-読点				7112474485684800			450			670	680	0	1		、	25875
OW core	OW6X 00000	680	690	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				1375251897600 0			460			680	690	0	1		ブンカ	50
OW core	OW6X 00000	690	710	I	芸術	ゲージュツ	芸術	芸術	芸術	名詞-普通名詞-一般				9222162401600000			470			690	710	0	1		ゲイジュツ	11482
OW core	OW6X 00000	710	720	I	に	ニ	に	に	に	助詞-格助詞				7745518285496832			480			710	720	和	1		ニ	28178
OW core	OW6X 00000	720	750	I	携わる	タズサワル	携わる	携わる	携わる	動詞-一般	五段-ワ行-一般	連体形-一般		6131985038844609			490			720	750	和	1		タズサワル	22308
OW core	OW6X 00000	750	770	I	人々	ヒトビト	人々	人々	人々	名詞-普通名詞-一般				8675163956716032			500			750	770	和	1		ヒトビト	31580
OW core	OW6X 00000	770	780	I	に	ニ	に	に	に	助詞-格助詞				7745518285496832			510			770	780	和	1		ニ	28178
OW core	OW6X 00000	780	790	I	一定	イッテー	一定	一定	一定	名詞-普通名詞-サ変形状詞可能				1375251897600 0			520			780	790	和	1		イッテー	2242
OW core	OW6X 00000	790	810	I	範囲	キカン	範囲	範囲	範囲	名詞-普通名詞-一般				6182948909972480			530			790	810	和	1		キカン	6329
OW core	OW6X 00000	810	830	I	に	ニ	に	に	に	助詞-格助詞				2289846211056537 6			540			810	830	和	1		ニ	33
OW core	OW6X 00000	830	840	I	、	、	、	、	、	補助記号-読点				9079594552352			550			830	840	和	1		、	33550
OW core	OW6X 00000	840	860	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				9222162401600000			560			840	860	和	1		ブンカ	33550
OW core	OW6X 00000	860	880	I	交流	コーリュー	交流	交流	交流	名詞-普通名詞-サ変可能				3417290782756608			570			860	880	和	1		コウリュウ	12432
OW core	OW6X 00000	880	890	I	促	ソク	促	促	促	接尾辞-名詞的-一般				4103111140581888			580			880	890	和	1		ソク	14927
OW core	OW6X 00000	890	900	I	と	ト	と	と	と	補助記号-括弧開				9235447246848 96			590			890	900	和	1		ト	34
OW core	OW6X 00000	900	910	I	と	ト	と	と	と	助詞-格助詞				7109014038299136			600			900	910	和	1		ト	25826
OW core	OW6X 00000	910	920	I	ど	ド	ど	ど	ど	動詞-非自立可能	サ行変格	連用形-一般		5370298291593857			610			910	920	和	1		ドル	19937
OW core	OW6X 00000	920	930	I	で	デ	で	で	で	助詞-格助詞				6837321680953856			620			920	930	和	1		デ	24814
OW core	OW6X 00000	930	950	I	世界	セカイ	世界	世界	世界	名詞-普通名詞-一般				5512401016948376			630			930	950	和	1		セカイ	20054
OW core	OW6X 00000	950	960	I	の	ノ	の	の	の	助詞-格助詞				7980444286028416			640			950	960	和	1		ノ	28989
OW core	OW6X 00000	960	980	I	人々	ヒトビト	人々	人々	人々	名詞-普通名詞-一般				8675163956716032			650			960	980	和	1		ヒトビト	31580
OW core	OW6X 00000	980	990	I	の	ノ	の	の	の	助詞-格助詞				7980444286028416			660			980	990	和	1		ノ	28989
OW core	OW6X 00000	990	1010	I	日本	ニッポン	日本	日本	日本	名詞-固有名詞-地名-国				7821659499274752			670			990	1010	固	1		ニッポン	28455
OW core	OW6X 00000	1010	1030	I	文化	ブンカ	文化	文化	文化	名詞-普通名詞-一般				9222162401600000			680			1010	1030	連	1		ブンカ	33550
OW core	OW6X 00000	1030	1040	I	へ	エ	へ	へ	へ	助詞-格助詞				9296104555856736			690			1030	1040	和	1		ヘ	33819
OW core	OW6X 00000	1040	1050	I	の	ノ	の	の	の	助詞-格助詞				7980444286028416			700			1040	1050	和	1		ノ	28989
OW core	OW6X 00000	1050	1070	I	理解	リカイ	理解	理解	理解	名詞-普通名詞-サ変可能				1109651632095319 04			710			1050	1070	連	0		リカイ	38891

⑨ 文字テーブル

サンプルID	文字開始位置	文字終了位置	文字	固定長フラグ	可変長フラグ
OW6X_00000	10	20	1	0	1
OW6X_00000	20	30		0	1
OW6X_00000	30	40	日	0	1
OW6X_00000	40	50	本	0	1
OW6X_00000	50	60	文	0	1
OW6X_00000	60	70	化	0	1
OW6X_00000	70	80	の	0	1
OW6X_00000	80	90	発	0	1
OW6X_00000	90	100	信	0	1
OW6X_00000	100	110	に	0	1
OW6X_00000	110	120	よ	0	1
OW6X_00000	120	130	る	0	1
OW6X_00000	130	140	国	0	1
OW6X_00000	140	150	際	0	1
OW6X_00000	150	160	文	0	1
OW6X_00000	160	170	化	0	1
OW6X_00000	170	180	交	0	1
OW6X_00000	180	190	流	0	1
OW6X_00000	190	200	の	0	1
OW6X_00000	200	210	推	0	1
OW6X_00000	210	220	進	0	1
OW6X_00000	220	230	(	0	1
OW6X_00000	230	240	1	0	1
OW6X_00000	240	250	)	0	1
OW6X_00000	250	260	文	0	1
OW6X_00000	260	270	化	0	1
OW6X_00000	270	280	庁	0	1
OW6X_00000	280	290	文	0	1
OW6X_00000	290	300	化	0	1
OW6X_00000	300	310	交	0	1
OW6X_00000	310	320	流	0	1
OW6X_00000	320	330	使	0	1
OW6X_00000	330	340	事	0	1
OW6X_00000	340	350	業	0	1
OW6X_00000	350	360	1	0	1
OW6X_00000	360	370		0	1
OW6X_00000	370	380	文	0	1
OW6X_00000	380	390	化	0	1
OW6X_00000	390	400	庁	0	1
OW6X_00000	400	410	文	0	1
OW6X_00000	410	420	化	0	1
OW6X_00000	420	430	交	0	1
OW6X_00000	430	440	流	0	1
OW6X_00000	440	450	使	0	1
OW6X_00000	450	460	事	0	1
OW6X_00000	460	470	業	0	1
OW6X_00000	470	480		0	1
OW6X_00000	480	490	文	0	1
OW6X_00000	490	500	化	0	1
OW6X_00000	500	510	庁	0	1
OW6X_00000	510	520	文	0	1
OW6X_00000	520	530	化	0	1
OW6X_00000	530	540	交	0	1
OW6X_00000	540	550	流	0	1
OW6X_00000	550	560	使	0	1
OW6X_00000	560	570	事	0	1
OW6X_00000	570	580	業	0	1
OW6X_00000	580	590	は	0	1
OW6X_00000	590	600	、	0	1
OW6X_00000	600	610	芸	0	1
OW6X_00000	610	620	術	0	1
OW6X_00000	620	630	家	0	1
OW6X_00000	630	640	、	0	1
OW6X_00000	640	650	文	0	1
OW6X_00000	650	660	化	0	1
OW6X_00000	660	670	人	0	1
OW6X_00000	670	680	等	0	1
OW6X_00000	680	690	、	0	1
OW6X_00000	690	700	文	0	1
OW6X_00000	700	710	化	0	1
OW6X_00000	710	720	に	0	1
OW6X_00000	720	730	携	0	1
OW6X_00000	730	740	わ	0	1
OW6X_00000	740	750	る	0	1
OW6X_00000	750	760	人	0	1
OW6X_00000	760	770	々	0	1
OW6X_00000	770	780	に	0	1
OW6X_00000	780	790	、	0	1
OW6X_00000	790	800	一	0	1
OW6X_00000	800	810	定	0	1
OW6X_00000	810	820	期	0	1
OW6X_00000	820	830	間	0	1
OW6X_00000	830	840	「	0	1
OW6X_00000	840	850	文	0	1
OW6X_00000	850	860	化	0	1
OW6X_00000	860	870	交	0	1
OW6X_00000	870	880	流	0	1
OW6X_00000	880	890	使	0	1
OW6X_00000	890	900	」	0	1
OW6X_00000	900	910	と	0	1
OW6X_00000	910	920	し	0	1
OW6X_00000	920	930	て	0	1
OW6X_00000	930	940	世	0	1
OW6X_00000	940	950	界	0	1
OW6X_00000	950	960	の	0	1
OW6X_00000	960	970	人	0	1
OW6X_00000	970	980	々	0	1
OW6X_00000	980	990	の	0	1
OW6X_00000	990	1000	日	0	1
OW6X_00000	1000	1010	本	0	1

⑩ 文字修正テーブル

サンプルID	文字開始位置	文字終了位置	修正型	原文文字	更新日時	更新作業者	メモ
OW6X_00007	3760	3770	erratum	他	2008/8/7 2:32	user1	
OW6X_00008	63451	63470	erratum	工	2008/8/13 6:49	user1	工→行
OW6X_00008	78860	78870	erratum	は	2008/4/15 16:35	user4	は→な
OW6X_00008	80830	80840	erratum	は	2008/4/15 16:35	user4	は→な
OW6X_00008	85331	85350	erratum	工	2008/4/15 16:34	user4	工→行
OW6X_00010	6482	6500	omission		2008/8/13 6:50	user1	脱字
OW6X_00010	27011	27030	erratum	工	2008/4/15 16:34	user4	工→行
OW6X_00012	2240	2240	excess	を	2008/8/13 6:50	user1	衍字
OW6X_00012	20170	20171	erratum	避	2008/5/23 15:41	user9	電子化誤り
OW6X_00014	9611	9630	erratum	会	2008/8/13 6:50	user1	電子化ママ
OW6X_00014	14120	14121	omission		2008/9/17 11:46	user8	を(脱字)
OW6X_00014	26720	26740	erratum	接	2008/8/4 10:08	user9	電子化誤り
OW6X_00016	47290	47291	erratum	範	2008/8/8 2:36	user1	範囲→規範
OW6X_00016	47291	47310	erratum	囲	2008/5/30 11:52	user4	範囲→規範
OW6X_00016	69401	69420	erratum	員	2008/5/30 12:06	user4	要員→要因
OW6X_00016	75350	75351	erratum	用	2008/5/28 15:04	user4	用意→容易
OW6X_00016	75351	75370	erratum	意	2008/5/28 15:04	user4	用意→容易

⑪ 数字テーブル

サンプルID	文字開始位置	文字終了位置	出現書字形	変換型	原文文字列
OW6X_00000	1600	1620	十五	decimal	15
OW6X_00000	3810	3830	十六	decimal	16
OW6X_00000	4020	4040	十一	decimal	11
OW6X_00000	5610	5630	十五	decimal	15
OW6X_00000	5660	5680	十六	decimal	16
OW6X_00000	5940	5960	十五	decimal	15
OW6X_00000	9420	9440	十五	decimal	15
OW6X_00000	9600	9620	十六	decimal	16
OW6X_00000	9650	9670	十五	decimal	15
OW6X_00000	9750	9770	十一	decimal	11
OW6X_00000	10800	10820	十六	decimal	16
OW6X_00000	10850	10870	十六	decimal	16
OW6X_00000	11750	11770	十六	decimal	16
OW6X_00000	13630	13650	十三	decimal	13
OW6X_00000	13950	13970	十五	decimal	15
OW6X_00000	13980	14000	十	decimal	10
OW6X_00000	14030	14050	三十二	decimal	32
OW6X_00000	14150	14170	十七	decimal	17
OW6X_00000	14960	14980	十六	decimal	16
OW6X_00000	19320	19360	二百三十五万	decimal	235万
OW6X_00000	19400	19420	十五	decimal	15
OW6X_00000	19660	19680	十五	decimal	15
OW6X_00000	19690	19710	十一	decimal	11
OW6X_00000	21850	21870	十六	decimal	16
OW6X_00000	21920	21940	十二	decimal	12
OW6X_00000	22550	22570	十八	decimal	18
OW6X_00000	24050	24070	十六	decimal	16

⑫ 振り仮名テーブル

サンプルID	文字開始位置	文字終了位置	出現書字形	振り仮名
OW6X_00000	6530	6540		かく
OW6X_00000	6540	6550		しょう
OW6X_00000	7520	7530		かく
OW6X_00000	7530	7540		しょう
OW6X_00000	8840	8850		しょう
OW6X_00000	8850	8860		へい
OW6X_00000	14910	14920		かんが

【サンプルデータ】

⑬ タグテーブル

サンプルID	出現順	文字開始位置	文字終了位置	タグ
OW6X_00000	1	10	24190	<mergedSample />
OW6X_00000	2	10	10	<sample sampleID="OW6X_00000" version="20070814" type="variableLength" tagID="v000000" tagType="open"
OW6X_00000	3	10	10	<article articleID="OW6X_00000_V001" isWholeArticle="false" tagID="v000001" tagType="open"
OW6X_00000	4	10	10	<titleBlock tagID="v000002" tagType="open" />
OW6X_00000	5	10	10	<title tagID="v000003" tagType="open" />
OW6X_00000	6	10	220	<sentence type="quasi" />
OW6X_00000	7	220	220	<br type="automatic_original" />
OW6X_00000	8	220	220	<title tagID="v000003" tagType="close" />
OW6X_00000	9	220	220	<titleBlock tagID="v000002" tagType="close" />
OW6X_00000	10	220	220	<cluster tagID="v000008" tagType="open" />
OW6X_00000	11	220	220	<titleBlock tagID="v000007" tagType="open" />
OW6X_00000	12	220	220	<title tagID="v000008" tagType="open" />
OW6X_00000	13	220	350	<sentence type="quasi" />
OW6X_00000	14	350	350	<br type="automatic_original" />
OW6X_00000	15	350	350	<title tagID="v000008" tagType="close" />
OW6X_00000	16	350	350	<titleBlock tagID="v000007" tagType="close" />
OW6X_00000	17	350	350	<cluster tagID="v000011" tagType="open" />
OW6X_00000	18	350	350	<titleBlock tagID="v000012" tagType="open" />
OW6X_00000	19	350	350	<title tagID="v000013" tagType="open" />
OW6X_00000	20	350	470	<sentence type="quasi" />
OW6X_00000	21	350	360	<enclosedCharacter description="○" />
OW6X_00000	22	470	470	<br type="automatic_original" />
OW6X_00000	23	470	470	<title tagID="v000013" tagType="close" />
OW6X_00000	24	470	470	<titleBlock tagID="v000012" tagType="close" />
OW6X_00000	25	470	470	<paragraph tagID="v000017" tagType="open" />
OW6X_00000	26	470	1740	<sentence />
OW6X_00000	27	830	900	<quote />
OW6X_00000	28	1740	1740	<br type="automatic_original" />
OW6X_00000	29	1740	1740	<paragraph tagID="v000017" tagType="close" />
OW6X_00000	30	1740	1740	<paragraph tagID="v000021" tagType="open" />
OW6X_00000	31	1740	3780	<sentence />
OW6X_00000	32	1750	1820	<quote />
OW6X_00000	33	2370	2440	<quote />
OW6X_00000	34	2970	3050	<quote />
OW6X_00000	35	3580	3660	<quote />
OW6X_00000	36	3780	3780	<br type="automatic_original" />
OW6X_00000	37	3780	3780	<paragraph tagID="v000021" tagType="close" />
OW6X_00000	38	3780	3780	<paragraph tagID="v000028" tagType="open" />
OW6X_00000	39	3780	4520	<sentence />
OW6X_00000	40	3870	3940	<quote />
OW6X_00000	41	4060	4130	<quote />
OW6X_00000	42	4240	4320	<quote />
OW6X_00000	43	4520	5540	<sentence />
OW6X_00000	44	5540	5540	<br type="automatic_original" />
OW6X_00000	45	5540	5540	<paragraph tagID="v000028" tagType="close" />
OW6X_00000	46	5540	5540	<figureBlock tagID="v000035" tagType="open" />
OW6X_00000	47	5540	5540	<figure tagID="v000036" tagType="empty" />
OW6X_00000	48	5540	5540	<caption tagID="v000037" tagType="open" />
OW6X_00000	49	5540	5760	<sentence type="quasi" />
OW6X_00000	50	5760	5760	<br type="automatic_original" />
OW6X_00000	51	5760	5760	<caption tagID="v000037" tagType="close" />
OW6X_00000	52	5760	5760	<figureBlock tagID="v000035" tagType="close" />
OW6X_00000	53	5760	5760	<cluster tagID="v000011" tagType="close" />
OW6X_00000	54	5760	5760	<cluster tagID="v000040" tagType="open" />
OW6X_00000	55	5760	5760	<titleBlock tagID="v000041" tagType="open" />
OW6X_00000	56	5760	5760	<title tagID="v000042" tagType="open" />
OW6X_00000	57	5760	5910	<sentence type="quasi" />
OW6X_00000	58	5760	5770	<enclosedCharacter description="○" />
OW6X_00000	59	5910	5910	<br type="automatic_original" />
OW6X_00000	60	5910	5910	<title tagID="v000042" tagType="close" />
OW6X_00000	61	5910	5910	<titleBlock tagID="v000041" tagType="close" />
OW6X_00000	62	5910	5910	<paragraph tagID="v000046" tagType="open" />
OW6X_00000	63	5910	6490	<sentence />
OW6X_00000	64	6490	6490	<br type="automatic_original" />
OW6X_00000	65	6490	6490	<paragraph tagID="v000046" tagType="close" />
OW6X_00000	66	6490	6490	<paragraph tagID="v000049" tagType="open" />
OW6X_00000	67	6490	8150	<sentence />
OW6X_00000	68	6530	6540	<ruby rubyText="かく" />
OW6X_00000	69	6540	6550	<ruby rubyText="しょう" />
OW6X_00000	70	7520	7530	<ruby rubyText="かく" />
OW6X_00000	71	7530	7540	<ruby rubyText="しょう" />
OW6X_00000	72	7650	8030	<sentence />
OW6X_00000	73	8150	8150	<br type="automatic_original" />
OW6X_00000	74	8150	8150	<paragraph tagID="v000049" tagType="close" />
OW6X_00000	75	8150	8150	<figureBlock tagID="v000057" tagType="open" />
OW6X_00000	76	8150	8150	<figure tagID="v000058" tagType="empty" />
OW6X_00000	77	8150	8150	<caption tagID="v000059" tagType="open" />
OW6X_00000	78	8150	8340	<sentence type="quasi" />
OW6X_00000	79	8340	8340	<br type="automatic_original" />
OW6X_00000	80	8340	8340	<caption tagID="v000059" tagType="close" />
OW6X_00000	81	8340	8340	<figureBlock tagID="v000057" tagType="close" />
OW6X_00000	82	8340	8340	<cluster tagID="v000040" tagType="close" />
OW6X_00000	83	8340	8340	<cluster tagID="v000006" tagType="close" />
OW6X_00000	84	8340	8340	<cluster tagID="v000062" tagType="open" />
OW6X_00000	85	8340	8340	<titleBlock tagID="v000063" tagType="open" />
OW6X_00000	86	8340	8340	<title tagID="v000064" tagType="open" />
OW6X_00000	87	8340	8490	<sentence type="quasi" />
OW6X_00000	88	8490	8490	<br type="automatic_original" />
OW6X_00000	89	8490	8490	<title tagID="v000064" tagType="close" />
OW6X_00000	90	8490	8490	<titleBlock tagID="v000063" tagType="close" />
OW6X_00000	91	8490	8490	<paragraph tagID="v000067" tagType="open" />
OW6X_00000	92	8490	9570	<sentence />
OW6X_00000	93	8500	8610	<quote />
OW6X_00000	94	8840	8850	<ruby rubyText="しょう" />
OW6X_00000	95	8850	8860	<ruby rubyText="へい" />
OW6X_00000	96	9570	9570	<br type="automatic_original" />
OW6X_00000	97	9570	9570	<paragraph tagID="v000067" tagType="close" />
OW6X_00000	98	9570	9570	<paragraph tagID="v000073" tagType="open" />
OW6X_00000	99	9570	10730	<sentence />
OW6X_00000	100	9850	9930	<quote />

⑭ 長単位テーブル

サンプルID	長単位出現書字形	長単位品詞	長単位活用型	長単位活用形	長単位語彙素読み	長単位語彙素	長単位語彙素表記	タグ情報開始	タグ情報終了	メモ	書新作業者	書新日時	長単位開始位置	長単位終了位置	文字情報始位置	文字情報終了位置	範囲対応
OWBX_00000_	1	名詞-数詞			イチ	B	一						10	20	10	20	
OWBX_00000_	、	補助記号-読点				B	、						20	30	20	30	
OWBX_00000_	日本文化	名詞-普通名詞-一般			ニホンブンカ	B	日本文化						30	70	30	80	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						70	80	70	80	
OWBX_00000_	発信	名詞-普通名詞-サ変可能			ハッシン	B	発信						80	100	80	100	
OWBX_00000_	に	助詞-格助詞			ニ	B	に						100	110	100	110	
OWBX_00000_	よる	動詞-一般	五段-ラ行一般	連体形一般	ヨル	B	拠る						110	130	110	130	
OWBX_00000_	国際文化交流	名詞-普通名詞-サ変可能			コクサイブンカコウリュウ	B	国際文化交流						130	190	130	190	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						190	200	190	200	
OWBX_00000_	推進	名詞-普通名詞-サ変可能			スイシン	B	推進						200	220	200	220	
OWBX_00000_	（	補助記号-括弧開				B	（						220	230	220	230	
OWBX_00000_	1	名詞-数詞			イチ	B	一						230	240	230	240	
OWBX_00000_	）	補助記号-括弧閉				B	）						240	250	240	250	
OWBX_00000_	文化庁文化交流使事業	名詞-普通名詞-一般			ブンカチョウブンカコウリュウジギョウ	B	文化庁文化交流使事業						250	350	250	350	
OWBX_00000_	1	名詞-数詞			イチ	B	一	MARU	/MARU				350	360	350	360	
OWBX_00000_	、	補助記号-読点				B	、						360	370	360	370	
OWBX_00000_	文化庁文化交流使事業	名詞-普通名詞-一般			ブンカチョウブンカコウリュウジギョウ	B	文化庁文化交流使事業						370	470	370	470	
OWBX_00000_	、	補助記号-読点				B	、						470	480	470	480	
OWBX_00000_	文化庁文化交流使事業	名詞-普通名詞-一般			ブンカチョウブンカコウリュウジギョウ	B	文化庁文化交流使事業						480	580	480	580	
OWBX_00000_	は	助詞-係助詞			ハ	B	は						580	590	580	590	
OWBX_00000_	芸術家	名詞-普通名詞-一般			ゲイジュツカ	B	芸術家						590	600	590	600	
OWBX_00000_	、	補助記号-読点				B	、						600	630	600	630	
OWBX_00000_	文化人等	名詞-普通名詞-一般			ブンカジントウ	B	文化人等						630	640	630	640	
OWBX_00000_	の	補助記号-読点				B	の						640	660	640	660	
OWBX_00000_	交流	名詞-普通名詞-一般			ブンカ	B	交流						660	690	660	690	
OWBX_00000_	に	助詞-格助詞			ニ	B	に						690	710	690	710	
OWBX_00000_	携わる	動詞-一般	五段-ラ行一般	連体形一般	タズサワル	B	携わる						710	720	710	720	
OWBX_00000_	人々	名詞-普通名詞-一般			ヒトビト	B	人々						720	750	720	750	
OWBX_00000_	に	助詞-格助詞			ニ	B	に						750	770	750	770	
OWBX_00000_	一定期間	補助記号-括弧開				B	「						770	780	770	780	
OWBX_00000_	、	名詞-普通名詞-一般			イッテイキカン	B	一定期間						780	790	780	790	
OWBX_00000_	」	補助記号-括弧閉				B	」						790	830	790	830	
OWBX_00000_	文化交流使	名詞-普通名詞-一般			ブンカコウリュウシ	B	文化交流使						830	840	830	840	
OWBX_00000_	」	補助記号-括弧閉				B	」						840	890	840	890	
OWBX_00000_	として	助詞-格助詞			トシテ	B	として						890	900	890	900	
OWBX_00000_	世界	名詞-普通名詞-一般			セカイ	B	世界						900	930	900	930	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						930	950	930	950	
OWBX_00000_	人々	名詞-普通名詞-一般			ヒトビト	B	人々						950	980	950	980	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						980	990	980	990	
OWBX_00000_	日本文化	名詞-普通名詞-一般			ニホンブンカ	B	日本文化						990	1030	990	1030	
OWBX_00000_	へ	助詞-格助詞			ヘ	B	へ						1030	1040	1030	1040	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						1040	1050	1040	1050	
OWBX_00000_	理解	名詞-普通名詞-サ変可能			リカイ	B	理解						1050	1070	1050	1070	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						1070	1080	1070	1080	
OWBX_00000_	深化	名詞-普通名詞-サ変可能			シンカ	B	深化						1080	1100	1080	1100	
OWBX_00000_	や	補助記号-読点			ヤ	B	や						1100	1110	1100	1110	
OWBX_00000_	、	補助記号-読点				B	、						1110	1120	1110	1120	
OWBX_00000_	日本	名詞-固有名詞-地名一般			ニッポン	B	日本						1120	1140	1120	1140	
OWBX_00000_	と	助詞-格助詞			ト	B	と						1140	1150	1140	1150	
OWBX_00000_	外国	名詞-普通名詞-一般			ガイコク	B	外国						1150	1170	1150	1170	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						1170	1180	1170	1180	
OWBX_00000_	文化人	名詞-普通名詞-一般			ブンカジン	B	文化人						1180	1210	1180	1210	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						1210	1220	1210	1220	
OWBX_00000_	ネットワーク	名詞-普通名詞-一般			ネットワーク	B	ネットワーク						1220	1280	1220	1280	
OWBX_00000_	の	助詞-格助詞			ノ	B	の						1280	1290	1280	1290	
OWBX_00000_	形成・強化	名詞-普通名詞-サ変可能			ケイセイキョウカ	B	形成・強化						1290	1340	1290	1340	
OWBX_00000_	に	助詞-格助詞			ニ	B	に						1340	1350	1340	1350	
OWBX_00000_	つながる	動詞-一般	五段-ラ行一般	連体形一般	ツナガル	B	繋がる						1350	1390	1350	1390	
OWBX_00000_	活動	名詞-普通名詞-サ変可能			カツドウ	B	活動						1390	1410	1390	1410	
OWBX_00000_	を	助詞-格助詞			ヲ	B	を						1410	1420	1410	1420	
OWBX_00000_	展開し	動詞-一般	サ行変格	連用形一般	テンカイシ	B	展開する						1420	1450	1420	1450	
OWBX_00000_	ていもう	動詞-非自立可能	五段-ワア行一般	連体形一般	テモウ	B	ていもう						1450	1490	1450	1490	
OWBX_00000_	こと	名詞-普通名詞-一般			コト	B	事						1490	1510	1490	1510	
OWBX_00000_	を	助詞-格助詞			ヲ	B	を						1510	1520	1510	1520	
OWBX_00000_	目的	名詞-普通名詞-一般			モクテキ	B	目的						1520	1540	1520	1540	
OWBX_00000_	として	助詞-格助詞			トシテ	B	として						1540	1570	1540	1570	
OWBX_00000_	、	補助記号-読点				B	、						1570	1580	1570	1580	
OWBX_00000_	平成	名詞-固有名詞-一般			ヘイセイ	B	平成						1580	1600	1580	1600	
OWBX_00000_	十五年度	名詞-数詞			ジュウゴネンド	B	十五年度						1600	1640	1600	1640	
OWBX_00000_	から	助詞-格助詞			カラ	B	から						1640	1660	1640	1660	
OWBX_00000_	始め	動詞-非自立可能	下一段-マ行	連用形一般	ハジメル	B	始める						1660	1680	1680	1680	
OWBX_00000_	た	助動詞	助動詞-タ	連体形一般	タ	B	た						1680	1690	1680	1680	
OWBX_00000_	事業	名詞-普通名詞-一般			ジギョウ	B	事業						1690	1710	1680	1680	

【サンプルデータ】

⑮ 文節テーブル

サンプルID	文字開始位置	文字終了位置	文節1	文節2	更新作業者	更新日時
OW6X_00000	10	0	B	*		
OW6X_00000	30	0	B	*	user30	2008/12/8 17:55
OW6X_00000	80	0	B	*		
OW6X_00000	130	0	B	*		
OW6X_00000	200	0	B	*		
OW6X_00000	220	0	B	*		
OW6X_00000	250	0	B	*		
OW6X_00000	350	0	B	*		
OW6X_00000	370	0	B	*		
OW6X_00000	470	0	B	*		
OW6X_00000	600	0	B	*		
OW6X_00000	690	0	B	*		
OW6X_00000	720	0	B	*		
OW6X_00000	750	0	B	*		
OW6X_00000	790	0	B	*		
OW6X_00000	830	0	B	*		
OW6X_00000	930	0	B	*		
OW6X_00000	960	0	B	*		
OW6X_00000	990	0	B	*		
OW6X_00000	1050	0	B	*		
OW6X_00000	1080	0	B	*		
OW6X_00000	1120	0	B	*		
OW6X_00000	1150	0	B	*		
OW6X_00000	1180	0	B	*		
OW6X_00000	1220	0	B	*		
OW6X_00000	1290	0	B	*		

⑯ 長単位語彙表テーブル

長単位出現書字形	長単位品詞	長単位活用型	長単位活用形	長単位語彙素読み	長単位語彙素	ID
日本らしい	接尾辞-形容詞的	形容詞	連体形-一般	ニホンラシイ	日本らしい	46102
日本人離れし	動詞-一般	サ行変格	連用形-一般	ニホンジンハナレスル	日本人離れ為る	46194
ニッポン	名詞-固有名詞-地名-国			ニッポン	日本	10706
ニホン	名詞-固有名詞-地名-国			ニッポン	日本	10715
日本	名詞-固有名詞-地名-国			ニッポン	日本	46099
日本以外	名詞-普通名詞-副詞可能			ニッポンイガイ	日本以外	46196
日本側専門家	名詞-普通名詞-一般			ニッポンガワセンモンカ	日本側専門家	46203
日本企業	名詞-普通名詞-一般			ニッポンキギョウ	日本企業	46197
日本教科書正常化運動本部	名詞-普通名詞-一般			ニッポンキョウカショセイジョウカウンドウホンブ	日本教科書正常化運動本部	46281
日本経済	名詞-普通名詞-一般			ニッポンケイザイ	日本経済	46338
日本国際博覧会	名詞-普通名詞-一般			ニッポンコクサイハクランカイ	日本国際博覧会	46238
日本国民	名詞-普通名詞-一般			ニッポンコクミン	日本国民	46234
日本時間	名詞-普通名詞-副詞可能			ニッポンジカン	日本時間	78662
日本中	名詞-固有名詞-地名-国			ニッポンジュウ	日本中	46159
日本人	名詞-普通名詞-一般			ニッポンジン	日本人	46161
日本人像	名詞-普通名詞-一般			ニッポンジンゾウ	日本人像	46169
日本信販	名詞-普通名詞-一般			ニッポンシンパン	日本信販	46200
日本人拉致事件	名詞-普通名詞-一般			ニッポンジンラチジケン	日本人拉致事件	46178
日本政府	名詞-固有名詞-組織名			ニッポンセイフ	日本政府	46273
日本政府	名詞-普通名詞-一般			ニッポンセイフ	日本政府	46274
日本全体	名詞-普通名詞-一般			ニッポンゼンタイ	日本全体	46204
日本チーム	名詞-普通名詞-一般			ニッポンチーム	日本チーム	46120
日本テレコム株	名詞-普通名詞-一般			ニッポンテレコムカブ	日本テレコム株	46124
日本テレビ	名詞-普通名詞-一般			ニッポンテレビ	日本テレビ	46125
日本特殊陶業	名詞-普通名詞-一般			ニッポントクシュトウギョウ	日本特殊陶業	46319
日本版スペースシャトル	名詞-普通名詞-一般			ニッポンバンスペースシャトル	日本版スペースシャトル	46318
日本部	名詞-普通名詞-一般			ニッポンブ	日本部	46379
日本フライングディスク協会副会長	名詞-普通名詞-一般			ニッポンフライングディスクキョウカイフクカイチョウ	日本フライングディスク協会副会長	46143
日本舗道	名詞-普通名詞-一般			ニッポンホドウ	日本舗道	46387
日本アジア航空	名詞-普通名詞-一般			ニホンアジアコウクウ	日本アジア航空	46103
日本ASEAN交流年	名詞-普通名詞-一般			ニホンアセアンコウリュウネン	日本アセアン交流年	46399
日本アマチュア選手権	名詞-普通名詞-一般			ニホンアマチュアセンシュケン	日本アマチュア選手権	46104
日本育英会	名詞-普通名詞-一般			ニホンイクエイカイ	日本育英会	46345
日本育英会奨学金貸与人員総数	名詞-普通名詞-一般			ニホンイクエイカイショウガクキンタイヨジンインソウス	日本育英会奨学金貸与人員総数	46346
日本医師会	名詞-普通名詞-一般			ニホンイシカイ	日本医師会	46217
日本―イタリア代表戦	名詞-普通名詞-一般			ニホンイタリアダイヒョウセン	日本―イタリア代表戦	46100
日本一	名詞-数詞			ニホンイチ	日本一	46157
日本一軍団	名詞-普通名詞-一般			ニホンイチグンダン	日本一軍団	46158
日本受入れ先	名詞-普通名詞-副詞可能			ニホンウケイレサキ	日本受け入れ先	46222
日本鰻輸入組合	名詞-普通名詞-一般			ニホンウナギユニュウクミアイ	日本鰻輸入組合	46398
日本映画	名詞-普通名詞-一般			ニホンエイガ	日本映画	46291
日本映画界	名詞-普通名詞-一般			ニホンエイガカイ	日本映画界	46292
日本エネルギー経済研究所	名詞-普通名詞-一般			ニホンエネルギーケイザイケンキュウショ	日本エネルギー経済研究所	46105
日本円	名詞-普通名詞-一般			ニホンエン	日本円	46206
日本オーディオ協会主催	名詞-普通名詞-サ変可能			ニホンオーディオキョウカイシュサイ	日本オーディオ協会主催	46109
日本オプティカル	名詞-普通名詞-一般			ニホンオプティカル	日本オプティカル	46106
日本オプティカルマーケティング部	名詞-普通名詞-一般			ニホンオプティカルマーケティングブ	日本オプティカルマーケティング部	46107
日本オリンピック委員会	名詞-普通名詞-一般			ニホンオリンピックイインカイ	日本オリンピック委員会	46108
日本音楽著作権協会	名詞-普通名詞-一般			ニホンオンガクチョサクケンキョウカイ	日本音楽著作権協会	46392
日本画	名詞-普通名詞-一般			ニホンガ	日本画	46326
日本海	名詞-普通名詞-一般			ニホンカイ	日本海	46313
日本海軍	名詞-固有名詞-一般			ニホンカイグン	日本海軍	46314
日本外交	名詞-普通名詞-一般			ニホンガイコウ	日本外交	46243
日本家屋	名詞-普通名詞-一般			ニホンカオク	日本家屋	46255
日本画家	名詞-普通名詞-一般			ニホンガカ	日本画家	46327
日本化学産業	名詞-普通名詞-一般			ニホンカガクサンギョウ	日本化学産業	46216
日本学術振興会	名詞-普通名詞-一般			ニホンガクジュツシンコウカイ	日本学術振興会	46252
日本学術振興会特別研究員制度	名詞-普通名詞-一般			ニホンガクジュツシンコウカイトクベツケンキュウインセ	日本学術振興会特別研究員制度	46253
日本各地	名詞-普通名詞-一般			ニホンカクチ	日本各地	46225
日本型	名詞-普通名詞-一般			ニホンガタ	日本型	46240
日本型システム	名詞-普通名詞-一般			ニホンガタシステム	日本型システム	46241
日本学校農業クラブ北海道連盟	名詞-普通名詞-一般			ニホンガッコウノウギョウクラブホッカイドウレンメイ	日本学校農業クラブ北海道連盟	46251
日本株式会社	名詞-普通名詞-一般			ニホンカブシキカイシャ	日本株式会社	46300
日本髪	名詞-普通名詞-一般			ニホンカミ	日本髪	46396
ニホンカモシカ	名詞-普通名詞-一般			ニホンカモシカ	日本羚羊	10716
日本側	名詞-普通名詞-一般			ニホンガワ	日本側	46201
日本側出資比率	名詞-普通名詞-一般			ニホンガワシュッシヒリツ	日本側出資比率	46202
日本関連情報	名詞-普通名詞-一般			ニホンカンレンジョウホウ	日本関連情報	46388
日本企業	名詞-普通名詞-一般			ニホンキギョウ	日本企業	46198
日本技術者教育認定機構	名詞-普通名詞-一般			ニホンギジュツシャキョウイクニンテイキコウ	日本技術者教育認定機構	46268
日本球界	名詞-普通名詞-一般			ニホンキュウカイ	日本球界	46321
日本球界復帰	名詞-普通名詞-サ変可能			ニホンキュウカイフッキ	日本球界復帰	46322
日本鳥類学会	名詞-固有名詞-組織名			ニホンキョルイガッカイ	日本鳥類学会	46397

⑰ 分類語彙表テーブル

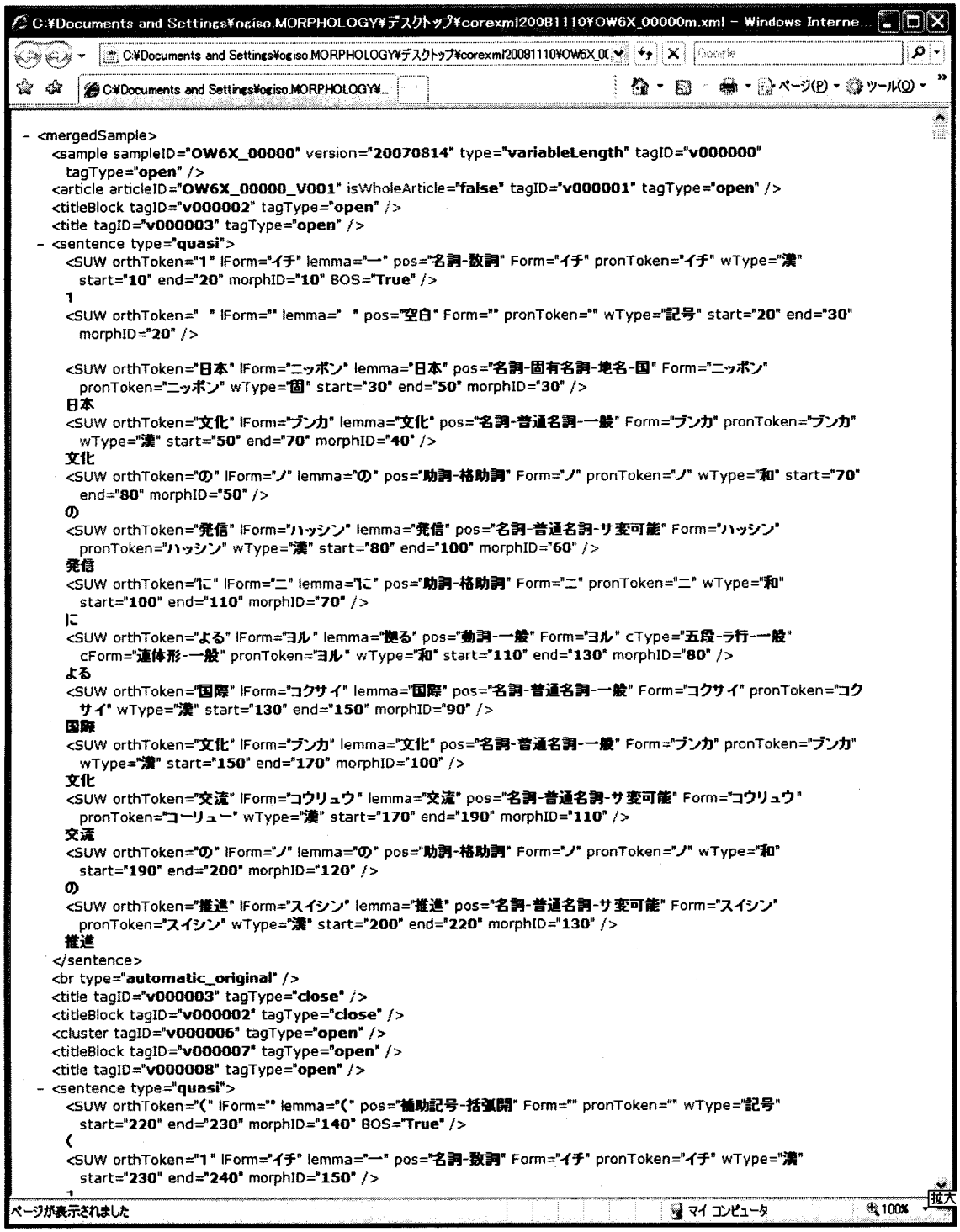
レコードID	見出し番号	レコード種別	部門	中項目	分類項目	分類番号	段落番号	小段落番号	語番	見出し	見出し本体	読み	逆読み	見出し本体 browu	読み:カタカナ	分類語彙表番号	更新作業者	更新日時	辞書装備	チェック済み	メモ
1066	1011A	体	用言	形容	属性-目的・成就	1.1113	4	1	1	目的	目的	もくてき	きてくも	目的	モクテキ	1.1113-04-01-01	user22	01 28 2009 8:04PM	0	1	
1835	1743A	体	用言	形容	成立	1.122	2	5	5	形成	形成	けいせい	いせいけ	形成	ケイセイ	1.1220-02-02-05	user22	01 29 2009 11:39AM	0	1	
3933	3724A	体	用言	作用	変化	1.15	16	5	5	変化	変化	しんか	かんし	変化	シンカ	1.1500-16-05-05	user22	01 30 2009 10:59AM	0	1	
3911	3706A	体	用言	作用	変化	1.15	16	1	1	変化	変化	きょうか	かうょき	変化	キョウカ	1.1500-16-01-01	user22	01 30 2009 10:57AM	0	1	
3826	3623A	体	用言	作用	変化	1.15	7	1	2	一定	一定	いってい	いてってい	一定	イッテイ	1.1500-07-01-02	user22	01 30 2009 10:52AM	0	1	
4457	4229A	体	用言	作用	動き	1.151	7	1	1	活動	活動	かつどう	うどつか	活動	カツドウ	1.1510-07-01-01	user22	01 30 2009 1:12PM	0	1	
5028	4780A	体	用言	作用	走り・飛び・流れなど	1.1522	9	3	2	交流	交流	こうりゅう	うりゅうこ	交流	コウリュウ	1.1522-09-03-02	user22	01 30 2009 3:30PM	0	1	
5425	5151A	体	用言	作用	突き・押し・引き・すれ違	1.1527	1	4	1	交流	交流	こうりゅう	うりゅうこ	交流	コウリュウ	1.1527-01-04-01	user22	01 30 2009 4:37PM	0	1	
6583	6258A	体	用言	作用	伸縮	1.1581	2	1	3	推進	推進	すいしん	んしいすい	推進	スイシン	1.1581-07-03-01	user22	02 2 2009 11:31AM	0	1	
7306	6955A	体	用言	作用	進歩・発達	1.1583	6	3	1	展開	展開	てんかい	いかんて	展開	テンカイ	1.1583-02-01-03	user22	02 2 2009 2:09PM	0	1	
7335	6984A	体	用言	作用	進歩・発達	1.1583	6	2	2	変化	変化	きょうか	かうょき	変化	キョウカ	1.1583-09-02-02	user22	02 2 2009 2:21PM	0	1	
7329	6978A	体	用言	作用	進歩・発達	1.1583	5	2	1	深化	深化	しんか	かんし	深化	シンカ	1.1583-05-02-02	user22	02 2 2009 2:30PM	0	1	
7907	7539A	体	用言	動作	動作	1.162	3	3	2	移動	移動	いどう	うどつか	移動	イドウ	1.1620-01-03-01	user22	02 2 2009 4:44PM	0	1	
9982	9187A	体	用言	変化	変化	1.1711	1	1	1	推進	推進	すいしん	んしいすい	推進	スイシン	1.1711-16-01-02	user22	02 3 2009 2:35PM	0	1	
10461	9920A	体	用言	変化	変化	1.1711	15	1	3	ネットワーク	ネットワーク	ねっとわーく	くわあとっね	ネットワーク	ネットワーク	1.1711-16-01-03	user22	02 3 2009 3:58PM	0	1	
14318	13867A	体	用言	数化	数化(一二三)	1.1951	15	1	3	第一(2-3)	第一	だいいち	ちいいだ	第一	ダイイチ	1.1951-15-01-03	user22	02 5 2009 4:36PM	0	1	
14316	13853A	体	用言	数化	数化(一二三)	1.196	3	1	3	百(Up2)	百	ひゃく	くゃひ	百	ヒャク	1.196-05-01-03	user22	02 5 2009 4:04PM	0	1	
14315	13853	体	用言	数化	数化(一二三)	1.196	3	2	3	十(Up3)	十	じゅう	うじゅ	十	ジュウ	1.196-05-01-02	user22	02 5 2009 4:03PM	0	1	
14330	13817	体	用言	数化	数化(一二三)	1.196	1	2	3	色(Up5)	色	いろ	ろい	色	イロ	1.196-01-02-03	user22	02 5 2009 3:57PM	0	1	
14329	13817	体	用言	数化	数化(一二三)	1.196	1	2	2	一(Up5)	一	いち	ちい	一	イチ	1.196-01-02-02	user22	02 5 2009 3:57PM	0	1	
15407	14530A	体	主体	人間	人間	1.2	6	1	3	二人(Up5)	二人	ふたり	りたふ	二人	フタリ	1.200-06-01-03	user22	02 6 2009 9:26AM	0	1	
15382	14512A	体	主体	人間	人間	1.2	4	1	1	人(ひとびと)	人	ひとびと	とびとひ	人	ジン	1.200-04-01-01	user22	01 5 2009 5:16PM	0	1	
21365	20282A	体	主体	公私	国	1.253	2	6	1	外国	外国	がいこく	くこいがい	外国	ガイコク	1.2530-02-06-01	user22	02 10 2009 3:40PM	0	1	
21421	20318A	体	主体	公私	国	1.253	4	3	3	国際	国際	こくさい	いさいこ	国際	コクサイ	1.2530-04-03-03	user22	02 10 2009 3:43PM	0	1	
21735	20827	1体	主体	公私	固有地名	1.259	1	2	日本(にほん)	日本	にっぽん	んぽんに	日本	ニッポン	1.2590-01-01-02	user22	02 10 2009 4:44PM	0	1		
21736	20827	1体	主体	公私	固有地名	1.259	1	3	日本(にっぽん)	日本	にっぽん	んぽんに	日本	ニッポン	1.2590-01-01-03	user22	02 10 2009 4:44PM	0	1		
22102	20988A	2体	主体	社会	社会・世界	1.26	4	1	1	世界	世界	せかい	いかせ	世界	セカイ	1.2600-04-01-01	user22	02 10 2009 6:02PM	0	1	
27368	26065A	体	活動	心身	注意・認知・了解	1.3062	13	2	1	理解	理解	りかい	いかり	理解	リカイ	1.3062-13-02-02	user22	02 13 2009 4:30PM	0	1	
32348	30885A	体	活動	表現	通信	1.3122	12	5	4	表情	表情	はっしん	んしっは	表情	ハッシン	1.3122-12-05-04	user22	02 17 2009 5:49PM	0	1	
32391	30928A	体	活動	表現	通信	1.3122	15	4	1	ネットワーク	ネットワーク	ねっとわーく	くわあとっね	ネットワーク	ネットワーク	1.3122-15-04-01	user22	02 17 2009 5:57PM	0	1	
35022	33458A	体	活動	生活	文化・風俗・風俗	1.322	1	1	1	芸術	芸術	げいじゅつ	つじゅいげ	芸術	ゲイジュツ	1.3220-01-01-01	user22	02 18 2009 6:08PM	0	1	
35763	34195A	体	活動	生活	文化・風俗・風俗	1.33	1	1	1	文化	文化	ぶんか	かんぶ	文化	ブンカ	1.3300-01-01-01	user22	02 19 2009 11:49AM	0	1	
40279	38584A	体	活動	行為・活動	行為・活動	1.343	15	3	1	活動	活動	かつどう	うどつか	活動	カツドウ	1.3430-15-03-01	user22	02 23 2009 11:35AM	0	1	
40873	38961A	体	活動	交わり	交わり	1.35	5	1	3	国際	国際	こくさい	いさいこ	国際	コクサイ	1.3500-05-01-03	user22	02 23 2009 2:03PM	0	1	
44914	43976A	体	活動	食事	料理・食事	1.38	4	1	1	食事	食事	しょくじ	じくょし	食事	ショクジ	1.3800-01-01-01	user22	02 25 2009 2:23PM	0	1	
44917	43980A	体	活動	食事	料理・食事・包装	1.386	3	1	1	形成	形成	けいせい	いせいけ	形成	ケイセイ	1.3860-01-03-01	user22	02 26 2009 2:23PM	0	1	
47787	45889A	体	活動	食事	料理・食事・包装	1.386	3	3	1	形成	形成	けいせい	いせいけ	形成	ケイセイ	1.3860-01-03-01	user22	02 4 2009 2:16PM	0	1	
47829	45456A	体	自然	自然	自然物	1.390	6	1	3	交流	交流	こうりゅう	うりゅうこ	交流	コウリュウ	1.3900-06-01-03	user22	02 4 2009 2:16PM	0	1	
63522	61190A	体	自然	自然	自然物・固有地名	1.3503	1	2	1	カメ	カメ	かめ	めか	カメ	カメ	1.3503-01-02-01	user22	01 28 2009 12:11PM	0	1	

⑱ 分類語彙表関連付けテーブル

分類語彙表番号	語彙素ID	更新作業者	更新日時
1.1000-03-01-01	12836	user20	2009/1/28 10:59
1.1113-04-01-01	37676	user20	2009/1/28 18:04
1.1220-02-02-05	10988	user20	2009/1/29 11:39
1.1500-07-01-02	2242	user20	2009/1/30 10:52
1.1500-16-01-01	9167	user20	2009/1/30 10:57
1.1500-16-05-05	17256	user20	2009/1/30 10:59
1.1510-07-01-01	6801	user20	2009/1/30 13:12
1.1522-09-03-02	12432	user20	2009/1/30 15:30
1.1527-01-04-01	12432	user20	2009/1/30 16:37
1.1562-02-01-03	18917	user20	2009/2/2 11:31
1.1581-07-03-01	25355	user20	2009/2/2 14:09
1.1583-02-01-03	25355	user20	2009/2/2 14:18
1.1583-05-02-02	17256	user20	2009/2/2 14:20
1.1583-06-02-02	9167	user20	2009/2/2 14:21
1.1620-01-03-01	8329	user20	2009/2/2 16:40
1.1660-07-02-03	18917	user20	2009/2/3 14:35
1.1711-18-01-02	28860	user20	2009/2/4 9:48
1.1951-15-01-03	25875	user20	2009/2/5 15:50
1.1960-01-02-02	2050	user20	2009/2/5 15:57
1.1960-01-02-03	2050	user20	2009/2/5 15:57
1.1960-05-01-02	18157	user20	2009/2/5 16:04
1.1960-05-01-03	18157	user20	2009/2/5 16:04
1.1961-04-01-01	2050	user20	2009/2/5 17:13
1.1962-03-01-02	18157	user20	2009/2/5 17:48
1.1962-03-01-03	18157	user20	2009/2/5 17:48
1.1962-10-03-02	25875	user20	2009/2/26 16:24
1.2000-04-01-01	31560	user20	2009/1/5 17:16
1.2000-06-01-03	18765	user20	2009/2/6 9:25
1.2000-06-02-03	5580	user20	2009/2/6 9:27
1.2530-02-06-01	7919	user20	2009/2/10 15:40
1.2530-04-03-03	12524	user20	2009/2/10 15:43
1.2590-01-01-02	28455	user20	2009/2/10 16:44
1.2590-01-01-03	28455	user20	2009/2/10 16:44
1.2600-04-01-01	20054	user20	2009/2/10 18:02
1.3062-13-02-02	39891	user20	2009/2/13 16:30
1.3122-12-05-04	29864	user20	2009/2/17 17:49
1.3122-15-04-01	28860	user20	2009/2/17 17:57
1.3220-01-01-01	11482	user20	2009/2/18 18:08
1.3300-01-01-01	33550	user20	2009/2/19 11:49
1.3430-15-03-01	6801	user20	2009/2/23 11:55
1.3500-05-01-03	12524	user20	2009/2/23 14:03
1.3800-04-01-01	17803	user20	2009/2/25 14:30
1.3860-01-03-01	10988	user20	2009/2/26 14:23
1.5001-06-01-03	12432	user20	2009/2/4 14:16
1.5503-01-02-01	7151	user20	2009/1/26 12:11

【サンプルデータ】

⑲ XML 形式のコアデータ



C:¥Documents and Settings¥ogiso.MORPHOLOGY¥デスクトップ¥corexml20081110¥OW6X_00000m.xml - Windows Interne...

C:¥Documents and Settings¥ogiso.MORPHOLOGY¥デスクトップ¥corexml20081110¥OW6X_00

C:¥Documents and Settings¥ogiso.MORPHOLOGY¥...

ページ(P) ツール(O)

```
- <mergedSample>
    <sample sampleID="OW6X_00000" version="20070814" type="variableLength" tagID="v000000"
      tagType="open" />
    <article articleID="OW6X_00000_V001" isWholeArticle="false" tagID="v000001" tagType="open" />
    <titleBlock tagID="v000002" tagType="open" />
    <title tagID="v000003" tagType="open" />
  - <sentence type="quasi">
      <SUW orthToken="1" lForm="イチ" lemma="一" pos="名詞-数詞" Form="イチ" pronToken="イチ" wType="漢"
        start="10" end="20" morphID="10" BOS="True" />
      1
      <SUW orthToken="　" lForm="" lemma="　" pos="空白" Form="" pronToken="" wType="記号" start="20" end="30"
        morphID="20" />

      <SUW orthToken="日本" lForm="ニッポン" lemma="日本" pos="名詞-固有名詞-地名-国" Form="ニッポン"
        pronToken="ニッポン" wType="固" start="30" end="50" morphID="30" />
      日本
      <SUW orthToken="文化" lForm="ブンカ" lemma="文化" pos="名詞-普通名詞-一般" Form="ブンカ" pronToken="ブンカ"
        wType="漢" start="50" end="70" morphID="40" />
      文化
      <SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-格助詞" Form="ノ" pronToken="ノ" wType="和" start="70"
        end="80" morphID="50" />
      の
      <SUW orthToken="発信" lForm="ハッシン" lemma="発信" pos="名詞-普通名詞-サ変可能" Form="ハッシン"
        pronToken="ハッシン" wType="漢" start="80" end="100" morphID="60" />
      発信
      <SUW orthToken="に" lForm="ニ" lemma="に" pos="助詞-格助詞" Form="ニ" pronToken="ニ" wType="和"
        start="100" end="110" morphID="70" />
      に
      <SUW orthToken="よる" lForm="ヨル" lemma="寄る" pos="動詞-一般" Form="ヨル" cType="五段-ラ行-一般"
        cForm="連体形-一般" pronToken="ヨル" wType="和" start="110" end="130" morphID="80" />
      よる
      <SUW orthToken="国際" lForm="コクサイ" lemma="国際" pos="名詞-普通名詞-一般" Form="コクサイ" pronToken="コク
        サイ" wType="漢" start="130" end="150" morphID="90" />
      国際
      <SUW orthToken="文化" lForm="ブンカ" lemma="文化" pos="名詞-普通名詞-一般" Form="ブンカ" pronToken="ブンカ"
        wType="漢" start="150" end="170" morphID="100" />
      文化
      <SUW orthToken="交流" lForm="コウリュウ" lemma="交流" pos="名詞-普通名詞-サ変可能" Form="コウリュウ"
        pronToken="コーリュー" wType="漢" start="170" end="190" morphID="110" />
      交流
      <SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-格助詞" Form="ノ" pronToken="ノ" wType="和"
        start="190" end="200" morphID="120" />
      の
      <SUW orthToken="推進" lForm="スイシン" lemma="推進" pos="名詞-普通名詞-サ変可能" Form="スイシン"
        pronToken="スイシン" wType="漢" start="200" end="220" morphID="130" />
      推進
    </sentence>
    <br type="automatic_original" />
    <title tagID="v000003" tagType="close" />
    <titleBlock tagID="v000002" tagType="close" />
    <cluster tagID="v000006" tagType="open" />
    <titleBlock tagID="v000007" tagType="open" />
    <title tagID="v000008" tagType="open" />
  - <sentence type="quasi">
      <SUW orthToken="(" lForm="" lemma="(" pos="補助記号-括弧開" Form="" pronToken="" wType="記号"
        start="220" end="230" morphID="140" BOS="True" />
      (
      <SUW orthToken="1" lForm="イチ" lemma="一" pos="名詞-数詞" Form="イチ" pronToken="イチ" wType="漢"
        start="230" end="240" morphID="150" />
      1
```

ページが表示されました　マイ コンピュータ　100%　拡大

図表目次

図

【図表目次】

表

研究開発部門言語資源グループ（形態論情報サブグループ）

小椋秀樹　（研究開発部門主任研究員）
小磯花絵　（研究開発部門研究員）
小木曽智信*　（研究開発部門研究員）
冨士池優美　（研究開発部門特別奨励研究員）
宮内佐夜香　（研究開発部門研究補佐員）
渡部涼子　（研究開発部門研究補佐員）
竹内ゆかり　（研究開発部門研究補佐員）
小川志乃　（研究開発部門研究補佐員）
小西光　（研究開発部門研究補佐員）
原裕　（研究開発部門非常勤研究員）
中村壮範*　（派遣社員，マンパワー・ジャパン株式会社）

（*印は執筆者）

国立国語研究所内部報告書（LR-CCG-08-04）
『現代日本語書き言葉均衡コーパス』
形態論情報データベースの設計と実装

平成２１年３月２４日

執 筆 者　小木曽智信　中村壮範

発 行 者　独立行政法人国立国語研究所

〒 190-8561　東京都立川市緑町 10 番地の 2

電話　042（540）4300（代表）

　（平 20-11）

国立国語研究所